

The international journal of science / 14 May 2020

outlook
COPD

nature



LINE OF SIGHT

Multichannel lidar
offers enhanced vision
for self-driving vehicles

Research integrity
Elisabeth Bik reveals
the secret life of an
image sleuth

School of rock
Archive of geological
samples needed to
unpick Earth's history

Coronavirus
Structural insight
into binding action
of SARS-CoV-2

Vol. 581, No. 7807
nature.com

Embed nature in strategies to reboot economies

The pandemic is devastating economies. As countries revive growth, they must recognize that recovery depends on ecological health.

The problem with the notion that nature is indestructible is this: it is wrong. Once economists accept that they are mistaken on this count, it could revolutionize the way in which we calculate economic progress – or the lack thereof – especially in developing countries.”

Thus wrote economist Partha Dasgupta in a *Nature* essay in 2008 (P. Dasgupta *Nature* 456, 44; 2008). The work is a lament for Earth, directed at his colleagues who study economic growth and mistakenly think that the natural resources on which we depend – from forests to fossil fuels – will always be there.

But he was optimistic, too. “It is only a matter of time before economics makes room for nature,” he wrote. And it seems he was right. Times have changed. More economics researchers are collaborating with ecologists – and vice versa. And businesses and government agencies have become ever more aware of our dependence on the fragile natural world. But there’s still one category of government department that is largely sticking to the old script. Most national ministries of finance and economics – arguably the most powerful departments of government – have been stubbornly resistant to redrawing their pictures of economic life. Until now.

The reports of the Millennium Ecosystem Assessment and the Intergovernmental Science–Policy Platform on Biodiversity and Ecosystem Services have helped to sensitize governments to biodiversity’s unprecedented rate of decline – and to the fact that at least one million species are at risk of extinction. And finance ministries have been taking note on what this means for their work.

For some years, policymakers across China have been adopting a metric called gross ecosystem product, which is a measure of the monetary value of those ecosystem goods and services that benefit people – such as flood protection or clean water.

Other countries are catching up, and last year, the UK Treasury turned to Dasgupta for advice, and gave him three ‘homework’ questions: what are biodiversity’s economic benefits; what are the economic costs when biodiversity is lost – for example, how much more would we have to pay if all pollination had to be performed by humans; and what practical actions can be taken to enhance both biodiversity and economic prosperity.

Last week, Dasgupta’s team released its work in progress

“As economies are revived, it is time to rebuild them in a way that takes nature’s true value into account.”

in the form of an interim report. It reminds its audience of economic policymakers that all human life is part of – and dependent on – natural capital and ecosystem services, which economic systems do not typically value, and which are now declining. If natural capital continues to shrink, standards of living will also eventually decline – even if growth, as measured by gross domestic product, continues to rise. Tantalizingly, the report holds back on providing examples of success stories and options for change, for which we must wait for the final version. The team is inviting feedback: readers can send comments until 1 June.

The interim report could not have been better timed – not only for the United Kingdom, but globally. The coronavirus pandemic is decimating economies. Heads of government, ministers of finance and lending agencies such as the World Bank and the International Monetary Fund are providing trillions of dollars in stimulus funding to keep economies going. But the need to urgently revive economic activity is fuelling concerns that this could come at the expense of environmental sustainability.

That isn’t the only worry. Several of this year’s key international environmental meetings – conferences led by the United Nations on climate change, biodiversity and the Sustainable Development Goals – have been postponed. The conference of the parties to the UN biodiversity convention, which would have taken place in Kunming, China, was meant to include a discussion on how and when to incorporate biodiversity into economic planning. That these meetings have been pushed back with no new scheduled dates should sound an alarm.

Fortunately, another key multilateral process – the next revision to the UN System of National Accounts (SNA) – is getting under way. The SNA is the global standard for measuring economic activity. Revising these rules is an infrequent event and will take five years to complete. Since the first publication in 1953, there have been just five revisions – the most recent in 2008. Now, for the first time, this process will include debate and discussion on how economic measurement can better account for sustainability and well-being – and also the value of the digital economy.

The significance of this development for environmental sustainability cannot be overstated. And it makes the Dasgupta review’s timing and message – that economic health depends on ecological health – even more important and necessary. It is also why the final report needs to be delivered on schedule this autumn, with concrete options for how nature can be incorporated into the work of institutions for economic planning.

Every nation’s economic plans and policies are rightly pivoting to dealing with COVID-19 and its effects. But as economies are revived, now is the right time to make up for past omissions – and rebuild them in a way that takes nature’s true value into account.

A strong signal from an independent report commissioned by the finance ministry of one of the world’s richest nations could go a long way in persuading its counterparts around the world that much-needed economic recovery must go with – and not against – the grain of nature.

Coronavirus drugs trials need scale and collaboration

The pandemic has given rise to an excess of small and uncontrolled clinical trials.

Researchers have rallied in unprecedented ways to defeat the coronavirus pandemic. They are retooling laboratories to focus on the virus; helping with testing efforts; and, in the case of clinician–researchers, working feverishly to carry out research studies while also treating patients in overwhelmed health-care systems.

Some clinical trials – such as the World Health Organization’s Solidarity trial of four potential COVID-19 therapies – are large and collaborative. They involve teams working together across many sites to test drug candidates against COVID-19. However, in the urgency to find treatments, other trials are smaller, do not always include a control group and don’t test medicines on enough patients to provide statistically meaningful results.

In the midst of a pandemic, there is a place for such initial exploration of potential treatments for those who are seriously ill. They can be quick to organize and do not need extensive resources – allowing clinicians in smaller hospitals and those with lower budgets to conduct research. But, in the end, the search for a successful drug needs the power of scale and the learning that comes from collaboration. More trials must, moreover, include control groups and ensure transparency with data.

Studies of the experimental antiviral drug remdesivir provide an example of the clinical chaos that can ensue when trials are not well designed. Remdesivir is widely considered to be among the best candidates for a drug against SARS-CoV-2, the virus that causes COVID-19. Over the past four months, a series of studies have been launched to investigate remdesivir’s effectiveness against COVID-19, but they have produced conflicting results.

Hopes were first raised by an early analysis of 53 people seriously ill with COVID-19 in the United States, Europe, Canada and Japan who were given remdesivir. Sixty-eight per cent showed a clinical improvement when given the drug (J. Grein *et al.* *N. Engl. J. Med.* <http://doi.org/ggrm99>; 2020). However, the study lacked a control group and was not an organized clinical trial – instead, it comprised observations of patients who had been given the drug in a last-ditch effort to save their lives.

By contrast, a randomized placebo-controlled trial of remdesivir conducted in China that started with 236 patients with COVID-19 found no significant benefit (Y. Wang *et al.* *Lancet* <http://doi.org/ggtgvt>; 2020). But enrolment in this trial was halted early when the outbreak

in China subsided, leaving the trial without enough participants to be able to detect relatively mild effects with statistical rigour.

Hopes were also raised when Gilead Sciences in Foster City, California – the company that makes remdesivir and holds the patent – released results on 29 April from a study of 397 people. It reported that patients can be treated as well with a five-day course of the drug as with a ten-day course, but because the study lacked a control group it was impossible to conclude with any certainty whether the drug had worked.

On the same day, the US National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, announced preliminary results from a randomized placebo-controlled trial with 1,063 participants. According to these preliminary results, those who received the drug were discharged from hospital or returned “to normal activity levels” after a median of 11 days in hospital, compared with 15 days for those given a placebo. But the results were announced at a press conference and the full data have not yet been released. So we do not know, for example, how often participants experienced unwanted side effects, or how well matched those in the control and treatment groups were in terms of age and other medical conditions.

Trials and tribulations

Two other medicines – hydroxychloroquine and chloroquine – provide another case study in the pitfalls of small and uncontrolled trials. After early studies in laboratory-grown cells suggested that the drugs might be effective against SARS-CoV-2 (M. Wang *et al.* *Cell Res.* **30**, 269–271; 2020), clinical trials were launched around the world. But in the wake of multiple trials – many of them small and uncontrolled – researchers still do not have a clear answer as to whether the drugs work against COVID-19 in people. Despite this – and despite their known effects on the heart – world leaders such as US President Donald Trump have fuelled a rush to take these drugs.

There is a different way. The REMAP-CAP study, for example, is a large study testing a variety of treatments against COVID-19, including hydroxychloroquine. It will include participants from more than 160 sites across 14 countries. The study takes advantage of sophisticated clinical-trial designs that allow researchers to add treatment groups to the trial as it is running – and to remove those that preliminary data indicate are not performing well. REMAP-CAP had the benefit of preparation time: it was originally designed to study pneumonia, and has since switched its focus to concentrate on COVID-19.

A pandemic emergency is a reason to work faster, but researchers must not lose sight of the fact that experimental interventions carry an inherent risk to the patient. To balance this risk, clinical trials must be as robustly designed as possible. Some trials need to be small, initial explorations of potential treatments; but, after that, researchers must think big. It’s important to move quickly to larger, collaborative trials – ones that span borders and share expertise – that have a greater chance of showing what really works.

“It’s important to move quickly to larger trials that have a greater chance of showing what really works.”

World view

Pandemic researchers — recruit your own best critics



By Daniël Lakens

To guard against rushed and sloppy science, build pressure testing into your research.

As researchers rush to find the best ways to quell the COVID-19 crisis, they want to get results out ultra-fast. Preprints — public but unvetted studies — are getting lots of attention. But even their advocates are seeing a problem. To keep up the speed of research and reduce sloppiness, scientists must find ways to build criticism into the process.

Finding ways to prove ourselves wrong is a scientific ideal, but it is rarely scientific practice. Openness to critique is nowhere near as widespread as researchers like to think. Scientists rarely implement procedures to receive and incorporate pushback. Most formal mechanisms are tied to the peer-review and publishing system. With preprints, the boldest peers will still criticize the work, but only after mistakes are made and, often, widely disseminated.

An initial version of a preprint by researchers at Stanford University in California estimated that COVID-19's fatality rate was 0.12–0.2% (E. Bendavid *et al.* Preprint at medRxiv <http://doi.org/dskd>; 2020). This low estimate was removed from a subsequent version, but it had already received widespread attention and news coverage. Many immediately pointed out flaws in how the sample was obtained and the statistics were calculated. Everyone would have benefited if the team had received this criticism before the data were collected and the results were shared.

It is time to adopt a 'red team' approach in science that integrates criticism into each step of the research process. A red team is a designated 'devil's advocate' charged with finding holes and errors in ongoing work and to challenge dominant assumptions, with the goal of improving project quality. The team has a role similar to that of 'white-hat hackers' hired in the software industry to identify security flaws before they can be discovered and exploited by malefactors. Similarly, teams of scientists should engage with red teams at each phase of a research project and incorporate their criticism. The logic is similar to the Registered Report publication system — in which protocols are reviewed before the results are known — except that criticism is not organized by journals. Ideally, there is a larger amount of speedier communication between researchers and their red team than peer review allows, resulting in higher-quality preprints and submissions for publication.

Even scientists who invite criticism from a red team acknowledge that it is difficult not to become defensive. The best time for scrutiny is before you have fallen in love with your results. And the more important the claims, the more scrutiny they deserve. The scientific process needs to incorporate methods to include 'severe' tests that will

Researchers need to commit to addressing criticism from the outset."

prove us wrong when we really are wrong.

An example of a large-scale collaboration that applies a red-team approach is the Psychological Science Accelerator (PSA), a global network of more than 500 psychology laboratories. The PSA has solicited research projects on questions related to the COVID-19 pandemic and has offered to assist with data collection. Projects range from effective risk communication to cognitive-reappraisal interventions. After researchers develop protocols, the PSA assembles a red team of experts in research ethics, measurement, data analysis and the project's field to offer criticism and to allow researchers to revise their protocols.

I reviewed one of these protocols after it had been submitted to a journal. I later saw the PSA reviews and learnt that I had repeated many criticisms, such as the generalizability of the stimulus and flexibility of the data analysis, that the red team had made — and that the researchers had opted to ignore.

This shows that assembling a red team isn't enough: research teams need to commit to addressing criticism from the outset. Sometimes, this is straightforward — items on checklists are absent from a proposal, or an independent statistical analysis yields different results, for example. Usually, it will be less clear whether criticism merits changing a protocol or including a caveat. The key is that, when results are presented, the team transparently communicates the criticism that the red team raised. (Perhaps incorporated criticism could be listed in the methods section of a paper, and unincorporated criticism in the limitations.) This will show how severely a claim has been tested.

Pushback on each step of a project should be recognized as valuable quality control and adherence to scientific values. Ideally, a research team could recruit its red team from group members not immediately involved in the project.

Incentives for red teams in science deserve special consideration. A red team might identify major flaws that mean a study should not proceed, so including a team member as a co-author on a future publication by the group would be a conflict of interest. In the computer-security industry, a red team is often paid if it uncovers serious errors. Computer scientist Donald Knuth famously gave out 'bug bounties' to people who uncovered technical errors in his published work. (Recipients often kept the small cheques as souvenirs, suggesting that social credit works as an incentive.) To investigate incentivized criticism, my group is now recruiting red-team members and offering financial rewards (see go.nature.com/3frpbjq).

With research moving faster than ever, scientists should invest in reducing their own bias and allowing others to transparently evaluate how much pushback their ideas have been subjected to. A scientific claim is as reliable as only the most severe criticism it has been able to withstand.

Daniël Lakens

is an associate professor in the human–technology interaction group at Eindhoven University of Technology, the Netherlands.
e-mail: d.lakens@tue.nl

News in brief



SCIENTISTS SUPPORT SCANDAL-HIT EPIDEMIOLOGIST

Scientists have rallied behind a UK government adviser on the coronavirus who resigned last week after revelations about his private life. Neil Ferguson, an influential epidemiologist at Imperial College London, left the UK government's Scientific Advisory Group for Emergencies (SAGE) following media reports that he contravened UK lockdown rules when a woman he was in a relationship with visited him. Ferguson's team did some of the modelling that led to the UK lockdown on 23 March. "I deeply regret any undermining of the clear messages around the continued need for social distancing," Ferguson told *The Daily Telegraph* newspaper.

Researchers say that, despite Ferguson's error, the incident highlights the scrutiny that scientists face during the pandemic. In an open letter, 26 scientists say that Ferguson's resignation has prompted efforts to discredit the scientific basis of policies such as lockdown. These, in turn, fuel a misconception that a single scientist is behind coronavirus policies, they say, whereas in reality many experts contribute.

Many lament Ferguson's departure from SAGE. "He shouldn't have had to resign," says Kieron Flanagan, a science-policy researcher at the University of Manchester, UK. The government has lost a very competent voice, he says.

US APPROVES FIRST CRISPR TEST FOR CORONAVIRUS

The US drug regulator has granted its first emergency-use approval for a new coronavirus test that takes advantage of the gene-editing technology CRISPR.

The US Food and Drug Administration's emergency-use authority allows it to make tests and drugs available faster than usual in a public-health emergency. The new diagnostic kit is based on an approach co-developed by CRISPR pioneer Feng Zhang at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. It will be used to test for the new coronavirus behind the ongoing pandemic, SARS-CoV-2, in laboratories that are certified to provide clinical tests.

Although the United States has ramped up testing in the past two weeks – averaging nearly 250,000 tests per day, according to the non-profit organization The COVID Tracking Project – there are test shortages in some places. Widespread use of the new kit approved by the Food and Drug Administration could help to alleviate backlogs and increase testing, says Mitchell O'Connell, a biochemist at the University of Rochester in New York, who was not involved in developing the test.

The kit has been developed by Sherlock Biosciences in Cambridge. It works by programming the CRISPR machinery, which has the ability to home in on certain genetic sequences, to detect a snippet of SARS-CoV-2 genetic material in a nose, mouth or throat swab, or in fluid from the lungs. If the virus's genetic material is found, a CRISPR enzyme generates a fluorescent glow. The test can return results in about an hour, according to the company.



CHINA IS PROMOTING UNPROVEN CORONAVIRUS TREATMENTS

The Chinese government is heavily promoting traditional medicines as treatments for COVID-19. The remedies, a major part of China's health-care system, are even being sent to countries including Iran and Italy as international aid. But scientists outside China say it is dangerous to support therapies that have yet to be proved safe and effective.

There are currently no proven treatments for the deadly respiratory disease caused by the new coronavirus, although many countries are trialling existing and experimental drugs. So far, only one treatment – the antiviral remdesivir – has been shown, in randomized control trials, to have some potential to speed up recovery.

In China, senior government officials and the state media are pushing a range of traditional Chinese medicines (TCMs) as being effective at alleviating symptoms and reducing deaths. However, there are no rigorous trial data to show that the remedies work.

Although the efficacy of some TCM remedies for COVID-19 is

being tested, some researchers say the trials have not been rigorously designed and are unlikely to produce reliable results. Government officials and TCM practitioners deem the remedies safe because many have been used for thousands of years, but some significant side effects have been reported.

"We are dealing with a serious infection which requires effective treatments. For TCM, there is no good evidence, and therefore its use is not just unjustified, but dangerous," says Edzard Ernst, a UK-based retired researcher into complementary medicines.

Other world leaders have promoted unproven treatments. US President Donald Trump has pushed the use of hydroxychloroquine, an antimalarial drug with significant potential side effects. And the president of Madagascar, Andry Rajoelina, has claimed that a herbal drink can cure people. But those leaders' claims have been criticized by scientists in their countries. By contrast, in China, criticism of TCM is muted.

News in focus



DAVID VAANKIN/THE WASHINGTON POST/GETTY

Children's susceptibility and immune response to the new coronavirus are hotly debated.

HOW DO CHILDREN SPREAD THE CORONAVIRUS? THE SCIENCE STILL ISN'T CLEAR

Schools are beginning to reopen – but scientists are still trying to understand what the deal is with kids and COVID-19.

By Smriti Mallapaty

The role of children in spreading the coronavirus has been a key question since the early days of the pandemic. Now, as some countries allow schools to begin reopening after weeks in lockdown, scientists are racing to figure it out.

Children represent a small fraction of confirmed COVID-19 cases – less than 2% of reported infections in China, Italy and the United States have been in people under 18 years old. But researchers are divided on whether children are less likely than adults to

get infected and to spread the virus.

Some say that a growing body of evidence suggests children are at lower risk. They are not responsible for the majority of transmission and the data support opening schools, says Alasdair Munro, a paediatric infectious-diseases researcher at University Hospital Southampton, UK. Children in Germany and Denmark have already returned to school, and students in some areas of Australia and France are set to go back gradually over the coming weeks.

Other scientists argue against a rushed return to classrooms. They say the incidence

of infection is lower in children than in adults partly because children haven't been exposed to the virus as much – especially with many schools closed. And children are not getting tested as often as adults, because they tend to have mild or no symptoms, the researchers say.

"I do not see any strong biological or epidemiological reason to believe that children don't get as infected," says Gary Wong, a researcher in paediatric respiratory medicine at the Chinese University of Hong Kong. "As long as there is community transmission in the adult population, reopening of schools will

likely facilitate transmission, as respiratory viruses are known to circulate in schools and day cares.” He says good surveillance and testing systems should be in place before schools reopen.

If children are driving the spread of the virus, infections will probably spike in the next few weeks in countries where children have already returned to school, say scientists.

But settling the debate will require large, high-quality population studies – some of which are already under way – that include tests for the presence of antibodies in the blood as a marker of previous infection.

Some scientists are studying children’s immune responses to find out why they have milder symptoms than adults when infected, and whether that offers clues to potential therapies.

Susceptibility debate

A study published on 27 April in *The Lancet Infectious Diseases*¹, which was first posted as a preprint in early March, analysed households with confirmed COVID-19 cases in Shenzhen, China. It found that children younger than ten years old were just as likely as adults to get infected, but less likely to have severe symptoms.

“That preprint really scared everybody,” says Munro, because it suggested that children could be silently spreading the infection.

But other studies, including some from South Korea, Italy and Iceland, where testing was more widespread, have observed lower infection rates among children. Some studies from China also support the suggestion that children are less susceptible to infection. One, published in *Science* on 29 April², analysed data from Hunan, where the contacts of people with known infections had been traced and tested for the virus. The authors found that for every infected child under the age of 15, there were close to 3 people infected between the ages of 20 and 64.

But the data are less conclusive for teenagers aged 15 years or older, and suggest that their risk of infection is similar to that of adults, says Munro.

Transmission risk

Even less well understood is whether infected children spread the virus in a similar way to adults. A study³ of a cluster of cases in the French Alps describes one nine-year-old who attended three schools and a skiing class while showing symptoms of COVID-19, but did not infect a single person. “It would be almost unheard of for an adult to be exposed to that many people and not infect anyone else,” says Munro.

Kirsty Short, a virologist at the University of Queensland in Brisbane, Australia, led an as-yet unpublished meta-analysis of several household studies, including some from

countries that had not closed schools at the time, such as Singapore. She found that children are rarely the first person to bring the infection into a home; they had the first identified case in only roughly 8% of households. By comparison, children had the first identified case during outbreaks of H5N1 avian influenza in some 50% of households, the study reports.

“The household studies are reassuring because even if there are a lot of infected children, they are not going home and infecting others,” says Munro.

But Wong argues that such research is biased, because the households weren’t randomly selected, but picked because there was already a known infected adult there. So it is

“Children have the least to gain from lockdowns, and they have a lot to lose.”

also very difficult to establish who introduced the virus, he says. School and day-care closures could also explain why children aren’t often the main source of infection with SARS-CoV-2. Other respiratory viruses can transmit from adults to children and back, so “I don’t believe this virus is an exception”, he says.

In fact, two preprints have reported that children with COVID-19 symptoms can have similar levels of viral RNA to adults. “Based on these results, we have to caution against an unlimited re-opening of schools and kindergartens in the present situation. Children may be as infectious as adults,” note the authors of one of the studies, led by Christian Drosten, a virologist at the Charité hospital in Berlin (see go.nature.com/2wccjps). However, it is not yet clear whether high levels of viral RNA are an indicator of how infectious a person is, notes Harish Nair, an epidemiologist at the University of Edinburgh, UK.

Transmission from schools to the broader community is not well studied, but an Australian report from an ongoing investigation suggests that such transmission is limited, and much lower than with other respiratory viruses, such as influenza (see go.nature.com/2yj88eq). The report looked at more than 850 people who had been in contact with 9 students and 9 staff members confirmed to have COVID-19 in primary and high schools in the state of New South Wales, and recorded only two cases of the disease among those contacts – both in children.

On the basis of the evidence, Munro says children should be allowed back to school. “Children have the least to gain from lockdowns, and they have a lot to lose,” such as missing out on education and not getting added social support such as free school meals, he says.

Schools reopening does not mean a return to normal, says Short. There will be lots of restrictions and changes, to reduce transmission risk, she says, such as moving desks apart in classrooms and closing playgrounds. Studies of transmission in schools as they reopen will also be important, says Wong. Researchers in the Netherlands plan to monitor this closely as schools open gradually over the coming weeks.

Immune response

Researchers do agree, however, that children tend to deal with COVID-19 better than adults. The majority of infected children have mild or no symptoms, but some do get very ill and even die. There have been reports of a small number of children in London and New York developing an inflammatory response similar to the rare childhood illness Kawasaki disease.

“I would not be surprised if COVID-19 is associated with Kawasaki disease, because many other viral infections have been associated with it,” says Wong. If the link proves to be genuine, it could have been missed in China, Japan and South Korea because Kawasaki disease is much more prevalent in Asia, he says.

One theory for why most children have milder symptoms, says Wong, is that children’s lungs might contain fewer or less-mature ACE2 receptors, proteins that the SARS-CoV-2 virus uses to enter cells. But to confirm this, researchers would need to study tissue samples from children, says Wong, and these are very difficult to get.

Others have suggested that children are more routinely exposed to other coronaviruses, such as those that cause the common cold, which protects them from serious disease. “But that doesn’t seem to hold much water, because even newborn babies don’t seem to get very severe disease” from the COVID-19 coronavirus, says Munro.

Wong suggests that children might mount a more appropriate immune response to the infection – strong enough to fight the virus, but not so strong that it causes major damage to their organs. He has done a preliminary analysis of 300 individuals infected with COVID-19 that has found that children produce much lower levels of cytokines – proteins released by the immune system. Patients of all ages with severe disease tend to have higher cytokine levels, he says. But he still needs to tease out the cause and effect. “Are they sicker because they have higher cytokine levels, or do they have higher cytokine levels because they are sicker?”

1. Bi, Q. et al. *Lancet Infect. Dis.* [https://doi.org/10.1016/S1473-3099\(20\)30287-5](https://doi.org/10.1016/S1473-3099(20)30287-5) (2020).

2. Zhang, J. et al. *Science* <https://doi.org/10.1126/science.abb8001> (2020).

3. Danis, K. et al. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/ciaa424> (2020).



COVID-19 testing of homeless people is rare, but necessary.

CORONAVIRUS SPREADS UNDER THE RADAR IN US HOMELESS SHELTERS

Controlling the spread of COVID-19 in group settings is essential to ending the outbreak, researchers say.

By Amy Maxmen

Researchers are beginning to test homeless individuals in the United States for the virus that causes COVID-19, and are discovering that the situation is out of control: tests are rare and outbreaks are spreading below the radar.

The lack of testing and assistance for people living in group settings – such as those in homeless shelters, nursing homes and prisons – threatens their lives as well as the nation's ability to curb COVID-19 because these communities can rapidly become the epicentres of new outbreaks that will spread, say researchers. Scientists are now scrambling to collect data and model the transmission of coronavirus under different group-living situations.

What they learn might protect not only the roughly 1.4 million people who use homeless shelters or transitional housing in the United States each year – a growing population as unemployment soars and prisons release people to ease crowding – but also other people who don't have the luxury of separating themselves from others. "What we're seeing in this first wave in the US is that the largest clusters are in populations where people don't have a lot of agency," says Gina Neff, a sociologist at the University of Oxford, UK.

"These populations will become the sources of new outbreaks, even when we feel like we kind of have it under control."

Shelters given space

Before COVID-19 was reported in China, Helen Chu, an infectious-disease specialist at the University of Washington in Seattle, and her colleagues were studying how the influenza virus spreads through homeless communities. "We wanted to develop a strategy that could be implemented for treatment and prevention in case a pandemic hit," she says. Coronavirus swooped in before they could finish. In March, Chu's team began surveying its study participants for the new coronavirus, too. So far, she says, most of those who have tested positive don't have obvious COVID-19 symptoms.

Researchers found something similar in Boston, Massachusetts. In one study, in which 147 people tested positive at one shelter, just 11 reported a cough and only 1 would have met the official criterion for testing – a fever (T. P. Baggett *et al.* *J. Am. Med. Assoc.* <http://doi.org/ggtsh3>; 2020). That study is changing practices at the network of shelters affiliated with the Boston Health Care for the Homeless Program, says Travis Baggett, director of research at the programme and an author on the study. "Our data show that if we aren't

more proactive, we'll be too late to prevent an outbreak," he says.

But most shelters still reserve tests for people with symptoms – or test broadly only after an outbreak has occurred. The results of this policy are troubling. For example, by the time a person from a shelter in San Francisco, California, had been diagnosed with COVID-19 in April, more than 90 other residents and 10 people who worked there were already infected. To influence policies, Baggett is running computer simulations to work out how many people will become infected, hospitalized or die from COVID-19 if the situation remains as it is – compared with the result if people are tested on a regular basis, regardless of symptoms. Costs are taken into account, too. "We're trying to inform policymakers about different ways of doing things," he says.

Towards a similar goal, a team of researchers from three US universities released a report in late March that lays out some minimal needs that might slow the spread of COVID-19 among homeless people, such as providing rooms for those at risk of severe disease because of underlying health conditions (see go.nature.com/3brFa5t). In projecting the "costs of inaction", they find that, without further interventions, more than 21,300 homeless people in the United States will need to be hospitalized for COVID-19, and 3,400 will die.

Canaries in the coal mine

Health departments in the United States have started implementing interventions, such as relocating homeless people to stadiums, where beds are spaced two metres apart. And in San Francisco, Seattle and other cities, officials have reserved hotels in which to isolate people with COVID-19 who don't have homes. Yet the vast majority of homeless individuals still remain in group facilities or in tents on the street, says Margot Kushel, a researcher-clinician who studies homelessness at the University of California, San Francisco.

She points out that many of the people sleeping in shelters have low-paid 'essential jobs', such as those in grocery shops and warehouses. This means they could become infected at work or in the shelters and spread the virus to others. Kushel says that, with data on how many people are infected in different settings, her team can estimate how often to screen, whether distributing face masks helps, and whether encampments are safer than indoor options. This last aspect matters in California, where about 91,000 people live outside.

But these calculations require much more data on rates of infection. The shortcoming is not necessarily because ample tests don't exist. For example, Shana McDevitt, a researcher involved with COVID-19 testing at the University of California, Berkeley, says that her team has extra testing capability, but doctors and health officials are reluctant to recommend

that everyone in a shelter is screened because officials lack plans for how to follow up on the results when infected people have no health insurance, money or housing. Furthermore, she says, a positive result means that the health department must work out who else the person might have had contact with – and screen them. It's a laborious task, but one McDevitt wants to see done. She says surveillance of homeless populations can also inform policy-makers about whether an outbreak is waxing or waning in their communities, because people there are so vulnerable to infections. "They're kind of a canary in the coal mine," she says.

Many social workers want a stronger public-health response, too. Donald Frazier, the executive director of Building Opportunities

for Self-Sufficiency, a non-profit organization based in Berkeley, says he can't let new individuals into his network's shelters without tests of their coronavirus status. A related problem, he says, is that California is releasing thousands of inmates from prisons to decrease the risk of outbreaks there, but they aren't being tested first – and many have nowhere to go.

Researchers working to dampen the toll of COVID-19 in other crowded spaces, such as nursing homes and meat-packing plants, worry that policymakers aren't concerned enough about outbreaks in marginalized populations. Kushel says, "As scientists, it's our role to raise up these issues and help the public understand how viruses do discriminate, since we live in an inequitable world."

potentially informing policy and speeding up research that could lead to the development of vaccines and treatments. But their popularity is spotlighting the scrutiny that these studies receive. Without peer review, it's hard to check the quality of the work, and sharing poor science could be harmful, especially when research can have immediate effects on medical practice. That has led platforms including bioRxiv and medRxiv to enhance their usual screening procedures.

"We've seen some crazy claims and predictions about things that might treat COVID-19," says Richard Sever, a co-founder of both servers.

Much of that speculative work has been based on computational models, says Sever – so, after consulting with experts in outbreak science, the team decided to bar those papers from bioRxiv. "We can't check the side effects of all the drugs and we're not going to peer-review to work out whether the modelling they're using has any basis," Sever says. "There are some things that should go through peer review, rather than being immediately disseminated as preprints."

Barabási understands the need to ensure patient safety but disagrees with the decision. "It's precisely the coronavirus that creates an environment where you need to share," he says. The purpose of a preprint server "is that we decide what is interesting, not the referees". He ended up posting the study on the physical-sciences preprint server arXiv.

Quality control

ArXiv, launched almost 30 years ago, was the first major preprint repository – but in recent years, discipline- and region-specific servers have mushroomed. Screening procedures vary, but an analysis of 44 servers, posted on 28 April on bioRxiv, found that most have quality-control systems (J.J. Kirkham *et al.* Preprint at bioRxiv <http://doi.org/dt3q>; 2020). Seventy-five per cent publicly provided information about their screening procedures, and 32% involved researchers in vetting articles for criteria such as relevance of content.

"There was perhaps a misconception that there are no screening checks that go on with preprint servers," says Jamie Kirkham, a biostatistician at the University of Manchester, UK, and a co-author of the study. "We have actually found that most of them do."

BioRxiv and medRxiv have a two-tiered vetting process. In the first stage, papers are examined by in-house staff who check for issues such as plagiarism and incompleteness. Then manuscripts are examined by volunteer academics or subject specialists who scan for non-scientific content and health or biosecurity risks. BioRxiv mainly uses principal investigators; medRxiv uses health professionals. Occasionally, screeners flag papers for further examination by Sever and members of the

HOW PREPRINT SERVERS ARE BLOCKING BAD CORONAVIRUS RESEARCH

Repositories have been flooded with studies – and are screening more closely to guard against poor science.

By Diana Kwon

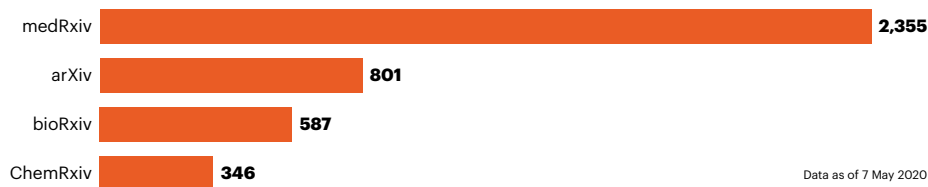
When Albert-László Barabási, a computational scientist at Northeastern University in Boston, Massachusetts, submitted a paper to the preprint server bioRxiv last month, he received an unexpected response. The biomedical repository would no longer accept manuscripts making predictions about treatments for COVID-19 solely on the basis of computational work. The bioRxiv team suggested that Barabási submit the study

to a journal for rapid peer review, instead of posting it as a preprint.

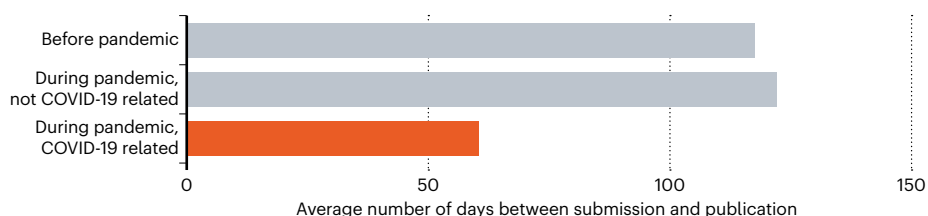
Publication norms are changing rapidly for science related to the coronavirus pandemic, as scientists worldwide conduct research at breakneck speed to tackle the crisis. Preprint servers – where scientists post manuscripts before peer review – have been flooded with studies. The two most popular for coronavirus research, bioRxiv and medRxiv, have posted some 3,000 studies on the topic (see 'Pandemic publishing'). The servers' merits are clear: results can be disseminated quickly,

PANDEMIC PUBLISHING

The major preprint servers have posted thousands of studies related to the coronavirus since the outbreak began.



Peer-reviewed journals have accelerated publication of studies on the coronavirus. One analysis of 14 titles, mainly in virology, found that the time to publish had dropped from 117 to 60 days.



leadership team. On bioRxiv, this is usually completed in 48 hours. On medRxiv, papers are scrutinized more closely because they may be more directly relevant to health, so the turnaround time is typically four to five days.

Sever emphasizes that the vetting process is mainly used to identify articles that might cause harm – for example, those claiming that vaccines cause autism or that smoking does not cause cancer – rather than to evaluate quality. For medical research, this also includes flagging papers that might contradict widely accepted public-health advice or inappropriately use causal language in reporting on a medical treatment.

But during the pandemic, screeners are watching for other types of content that need extra scrutiny – including papers that might fuel conspiracy theories. This extra screening

“There was perhaps a misconception that there are no screening checks that go on with preprint servers.”

was put in place at bioRxiv and medRxiv after a backlash against a now-withdrawn bioRxiv preprint that reported similarities between HIV and the new coronavirus, which scientists immediately criticized as poorly conducted science that would prop up a false narrative about the origin of SARS-CoV-2. “Normally, you don’t think of conspiracy theories as something that you should worry about,” Sever says.

These heightened checks and the sheer volume of submissions have meant that the servers have had to draft in more people. But even with the extra help, most bioRxiv and medRxiv staff have been working seven-day weeks, according to Sever. “The reality is that everybody’s working all the time.”

Growing trend

ArXiv and ChemRxiv, a preprint server for chemistry, have also seen their share of COVID-19 papers. ArXiv has posted more than 800 and ChemRxiv has around 200. Both platforms have enhanced their screening procedures for COVID-19-related papers, although neither has stopped posting all studies with treatment-related computational predictions. “If all the [preprint platforms] had the same standards, then we’d be systematically shutting out the same voices,” says Steinn Sigurdsson, arXiv’s scientific director. “We want to have somewhat overlapping domains.”

Marshall Brennan, ChemRxiv’s publishing manager, says that when it comes to papers about treatments, they are “taking much more liberty than we normally would to send those back to the authors to say, ‘Look, this science here is suitable for a preprint server, but you can’t make these claims in the context

of a public-health crisis.” He notes that, in one such paper, the authors had recommended a home remedy for COVID-19 entirely on the basis of a computational analysis. That paper was swiftly rejected.

Expedited publication

The abundance of coronavirus research is also reshaping peer review at journals. Several titles, including *Science*, journals published by Cell Press, *The BMJ* and *Nature*, report a surge in coronavirus-related submissions, and many have accelerated the peer-review process to ensure rapid dissemination.

A preprint posted in April on bioRxiv2 found that many medical-research journals had drastically speeded up publication pipelines for COVID-19 papers (S. P. J. M. Horbach. Preprint at bioRxiv <http://doi.org/dt3r;2020>). The analysis, which included 14 journals, found that average turnaround times had fallen from 117 to 60 days (see ‘Pandemic publishing’). (The study omitted several influential journals, such as *JAMA*, *The Lancet* and *The New England Journal of Medicine* because of a lack of appropriate data.) Some journals went from submission to publication in two weeks or less.

“That really makes one wonder how thorough this process really is,” says the study’s author, Serge Horbach, a doctoral student at Radboud University in Nijmegen, the Netherlands.

Howard Bauchner, the editor-in-chief of *JAMA*, notes that low-quality submissions are rising. Journals in the JAMA Network have received 53% more submissions in the first quarter of this year than in the same period in 2019. “Many of these are related to COVID-19, but most are of low quality,” Bauchner says.

To address the need for rapid review, a group of publishers and scholarly-communication organizations announced an initiative last month to accelerate the publication of COVID-19 papers using measures such as asking people with relevant expertise to join a list of rapid reviewers. The initiative’s members include Outbreak Science Rapid PREview, a platform where researchers can request or provide swift reviews of outbreak-related preprints.

Even in the light of expedited publication, it is important to remember that “the role of the journal is to say: ‘This has been fairly peer-reviewed, statistically reviewed, and can be relied on,’ rather than, ‘This is coming out at you as fast as it possibly can,’” says Theodora Bloom, executive editor of *The BMJ* and a co-founder of medRxiv. Still, Bloom notes that the COVID-19 papers submitted to her journal “are being handled at the fastest rate possible.”

Unlike preprint servers, being published in a journal gives papers the appearance of being reliable and valid knowledge, Horbach adds. “Nonsense or incorrect science in one of these papers is potentially much more harmful.”

Q&A

Pandemic economics



Economists are striving to make sense of the coronavirus pandemic’s dramatic effects on the economy. Arthur Turrell, a physicist-turned-researcher at the Bank of England, spoke to *Nature* about tracking the real-time and long-term financial impacts.

Has the pandemic changed your work?

It’s changed my focus. It’s boosted one of our efforts to provide better monitoring of the current economic situation for the bank’s policymakers. Typical macroeconomic data points, such as those on gross domestic product, come out quarterly. Now changes are happening weekly. And with policies such as lockdown, it’s like whole sectors of the economy have been turned off. So we’ve had to think differently. We’ve been using tools from data science and computer science to automatically collect and analyse data when they come out, and to create a report for policymakers.

What kind of research are you doing?

It’s important to understand the interaction between the macroeconomy and the progression of the disease. One project I’m working on is melding macroeconomic and epidemiological models. We slammed together two simple macroeconomic and epidemiological models. ‘Compartmental models’ in epidemiology study the dynamics of infectious diseases by dividing the population into groups, such as people who are infectious or recovered. It’s not that familiar to economists, but might be better known to those of us with science backgrounds. We’ve made most progress on that type of model for combining macroeconomics and epidemiology.

What can these models tell you?

For instance, perhaps people who have long-term health effects from the virus won’t go to work in the same way as before, or people will keep working from home. Those are economic impacts of the virus.

Interview by Elizabeth Gibney

This interview has been edited for length and clarity.



GABRIELA HASBUN FOR NATURE

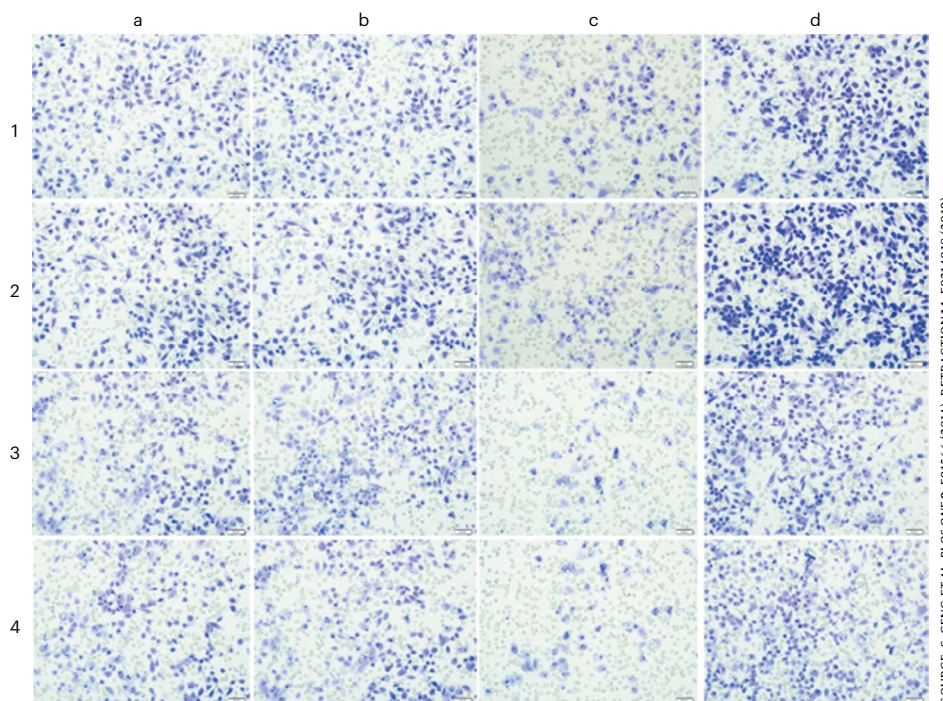
SEEING DOUBLE

Elisabeth Bik quit her job to spot errors in research papers — and has become the public face of image sleuthing. **By Helen Shen**



ARE YOU A SUPER-SPOTTER?

Elisabeth Bik identified five duplicated areas in these cell microscopy images from separate experiments in two figures in a paper. We have simplified the labels. Can you spot the duplications? (Answers overleaf).



SOURCE: S. CENG ET AL. PLOS ONE 9, E91566 (2014); RETRACTION 14, E0214018 (2019).

look at the Belgian message. Attached are images of western blots – the results of a common test to detect proteins in biological samples – from a published research paper. The writer wants to know: does Bik see anything fishy in this paper? Have these pictures been digitally altered?

Bik, a microbiologist from the Netherlands who moved to the United States almost two decades ago, is a widely lauded super-spotter of duplicated images in the scientific literature. On a typical day, she'll scan dozens of biomedical papers by eye, looking for instances in which images are reused and reported as results from different experiments, or where parts of images are cloned, flipped, shifted or rotated to create 'new' data (see 'Are you a super-spotter?'; answers overleaf at 'Did you spot them?').

Her skill and doggedness have earned her a worldwide following. "She has an uncommon ability to detect even the most complicated manipulation," says Enrico Bucci, co-founder of the research-integrity firm Resis in Samone, Italy. Not every issue means a paper is fraudulent or wrong. But some do, which causes deep concern for many researchers. "It's a terrible problem that we can't rely on some aspects of the scientific literature," says Ferric Fang, a microbiologist at the University of Washington, Seattle, who worked on a study with Bik in which she analysed more than 20,000 biomedical papers, finding problematic duplications in roughly 4% of them (E. M. Bik *et al.* *mBio* 7, e00809-16; 2016). "You have postdocs and students wasting months

or years chasing things which turn out to not be valid," he says.

Bik is not the world's only image sleuth, but she is unique in how publicly she presents her work. Many image checkers work behind the scenes, publishing their findings in research papers and writing privately to journals; a few are hired by journals or institutions. Some who flag up image problems work under pseudonyms, preferring not to be identified. But Bik posts her finds almost every day on Twitter and other online forums, in the process teaching others how to spot duplications and pressur-

"The things she calls out are usually real issues."

ing journals to investigate papers. In so doing, she's generated an "avalanche of reactions" and awareness about the problem, says Bucci. Bik estimates that her discoveries have led to at least 172 retractions and more than 300 errata and corrections – but all too often, she says, her warnings seem to be ignored.

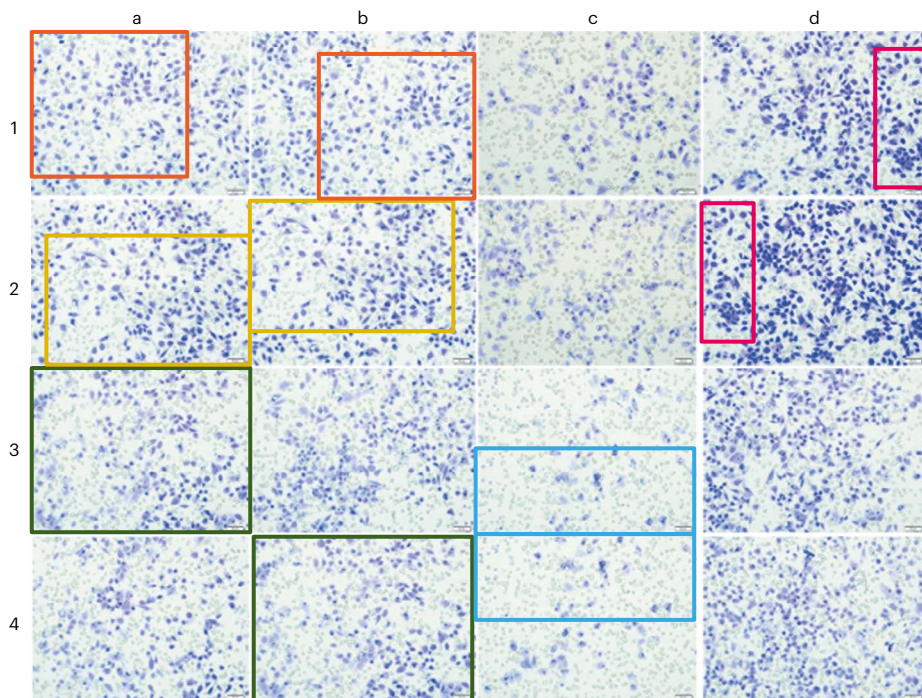
In April 2019, Bik announced that she had left her paid job at a biomedical start-up firm and would pursue image integrity work full-time, free of charge, for at least a year. A year on, she shows no signs of changing course – even though she has faced harassment, and at times been overwhelmed with requests. She's also shared her files with computer scientists trying to develop software to spot duplicated

February the fourteenth starts like most other days for Elisabeth Bik: checking her phone in bed, she scrolls through a slew of Twitter notifications and private messages from scientists seeking her detective services. Today's first request is from a researcher in Belgium: "Hi! I know you have a lot of people asking you to use your magic powers to analyse figures, blots and others but I just wanted to ask your opinion..."

After pouring a cup of coffee, Bik sits down at the long, wooden dining table that serves as her workstation at her home in Sunnyvale, California. She checks her e-mails on a giant 34-inch curved monitor, and takes a closer

DID YOU SPOT THEM?

Here are the duplicated sections Bik saw.



images across millions of papers, although the programs will probably always need human verification. “I’m enjoying it so much that I feel I just want to keep on doing this,” she says.

Hooked by a double smudge

Bik stumbled into image sleuthing around 2013, when, as a staff scientist at Stanford University in California, she read articles about scientific integrity and plagiarism. Out of curiosity, she googled quotes from her own published papers, and quickly found that other authors had lifted text without giving credit. “I was hooked. I was angry,” she says. “I immediately got fascinated about it, like how other people get fascinated by reading about crimes.” At one point, while examining a PhD thesis containing plagiarized text, something even more compelling caught her eye: a western-blot image with a distinctive smudge. The same image appeared in another chapter, supposedly for a different experiment. The chapters had also appeared as research articles, with the same errors, Bik saw. She e-mailed journal editors in January 2014; in June, she anonymously reported the papers online at PubPeer, a website where scientists can discuss published papers. These were Bik’s first reports of suspected manipulation in the literature. After an investigation by Case Western Reserve University in Ohio, the articles were retracted in 2015 and 2016.

Hunting for and cataloguing these images became a hobby. Then Bik contacted Fang and Arturo Casadevall, a microbiologist at Johns Hopkins University in Baltimore, Maryland. The trio decided that Bik’s rare talent could

lead to an in-depth inquiry of the frequency of problems in biomedical work. They sampled 20,621 papers, with Bik screening each – a task Fang says only she could do – before passing on her finds to Fang and Casadevall for corroboration. “It’s like a magic trick,” says Fang. “When it’s pointed out to you how it works, you can start to see it.” The team found 782 papers with what they termed “inappropriate” duplications, and Bik notified the relevant journals. The team reported the work in 2016 in *mBio*, at which Casadevall is editor-in-chief.

Bik spent so much of her spare time on duplicated images that last year she decided to leave her job as director of science at Astarte Medical in Foster City, California. “I realized I was getting more enthusiastic about image duplication work than my real job,” she says.

“It’s an impressive decision to make,” says Jennifer Byrne, a molecular biologist at the University of Sydney in Australia and herself a data-integrity sleuth who hunts for faulty genetic sequences in published papers. “It was very brave and, to be honest, pretty selfless.” Bik does not get paid for most of her work, but does some occasional paid consulting, and receives modest sums through a Patreon crowdfunding page. After decades of working and saving, she expects her current situation will be sustainable indefinitely.

The duplication database

Bik now operates out of a light-filled dining room, with floor-to-ceiling windows overlooking a garden filled with fruit trees and other plants, which she has catalogued in a spreadsheet. She also has a spreadsheet for

her collection of nearly 2,000 turtle figurines – gathered from travels and friends – which she keeps in a wall of glass cabinets. Most prized of all her spreadsheets, however, is a collection of more than 3,300 questionable papers, most of them flagged because of an issue with their images. (Bik sometimes raises other concerns with papers, such as around plagiarism or conflicts of interest.)

On a day without interruptions, Bik can peruse 100 papers or so, adding between 1 and 20 hits to her database (see ‘Advanced super-spotter test’; answers overleaf at ‘Did you spot them?’). A repeated smudge here or there, or a familiar smattering of data points: the visual indicators of duplication leap out at Bik from the screen. The collection is large enough to generate its own leads. It was looking at other papers by authors in her *mBio* data set, for instance, that led Bik last November to a case that generated her widest media coverage so far: a cluster of papers co-authored by Cao Xuetao, a prominent immunologist who has advocated for stronger research integrity in China, and who is the president of Nankai University in Tianjin. (Most of the articles listed Cao’s other affiliation, at the National Key Laboratory of Medical Immunology in Shanghai.) Bik and other pseudonymous commenters flagged apparent issues in more than 60 papers at PubPeer.

China’s ministry of education said it would investigate the articles, and Cao replied at PubPeer that he would re-examine the manuscripts, and that he was confident that the publications remained valid. Some authors replied swiftly on the site to point to honest errors. In one case, apparent duplicate images were in fact supposed to represent the same experiment but were not clearly labelled as such, an explanation that Bik accepts. In another, authors posted raw data and said the data seemed similar only after being processed for a paper. In still others, authors said there had been accidental mistakes, and by May this year, 13 of the flagged papers had received corrections, most stating that scientific conclusions weren’t affected. (Cao and China’s education ministry didn’t comment further for this article.)

Sometimes, Bik’s finds have pointed to suspected large-scale operations. This year, she and others have flagged a series of more than 400 papers that, they say, contain so many similarities that they could be the product of a ‘paper mill’ – a company that produces papers to order. Several image detectives worked to flag and collate the papers, including pseudonymous sleuths @mortenoxe, @TigerBB8 and @SmutClyde, who posted a list of papers in January, on a blog run by science journalist Leonid Schneider. “Finding these fabricated images should not rely solely on the work of unpaid volunteers,” Bik wrote in February on her own blog. Journals say they are now

SOURCE: S. GENGET AL PLOS ONE 9, E91566 (2014); RETRACTION 14, E0214018 (2019).

investigating the papers, many of which are authored by doctors in Chinese hospitals, and some retractions are already being prepared.

Bik's data have revealed some insights into factors that correlate with image duplication. Her *mBio* paper reported that duplicated images had a slight tendency to occur more frequently in lower-impact journals. The paper also examined a subset of 348 articles flagged in *PLoS ONE*: taking into account the frequency of publication in the journal, it seemed that papers from China and India were more likely to contain problematic images. But Bik doesn't target one country's authors, she says. "I search for problematic papers, regardless of what country they are from," she wrote in November. In all, Bik has flagged up duplications in papers with lead authors from 49 countries.

Gamified sleuthing

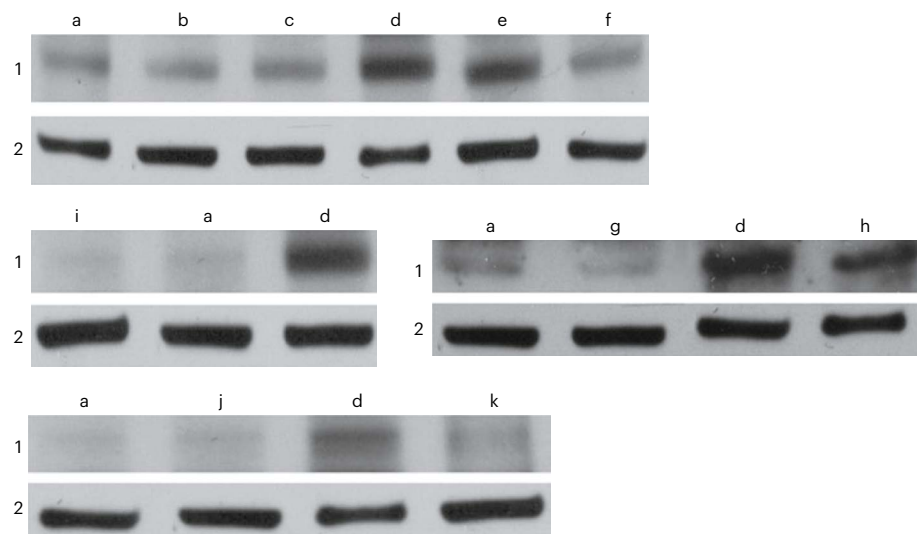
Nearly every day, Bik posts images with suspected problems to Twitter under the hashtag #ImageForensics, challenging her audience – which has almost tripled in the past year to more than 60,000 followers – to spot the matches before she posts her answers. The puzzles attract numerous guesses within minutes, and some eagle-eyed players spot issues that she misses. (She gives out emoji medals to top performers.) Bik says she hears from some followers who have picked up skills from her and spotted problematic images while peer-reviewing manuscripts. "I feel I'm changing people's way of looking at these images," she says. The work is sometimes overwhelming for Bik, who calls herself a "super introvert". Last November, she tweeted: "I am getting so many (anonymous) emails with people who want me to check certain authors or papers that I cannot possibly follow up. So many names ... And so much hidden pain among honest scientists about these dishonest coworkers."

Bik also posts detailed reports on what she sees to PubPeer, and occasionally comments there to support other tipsters. Many PubPeer users post their criticisms under pseudonyms – as does Bik in some cases, if she feels very worried about litigious authors. But she has posted more than 2,100 comments under her own name at the site since 2014. "What distinguishes Elisabeth is her willingness to identify herself, which is extremely admirable. It certainly helps with people taking the allegations seriously," says Mike Rossner, a former managing editor at the *Journal of Cell Biology* and president of Image Data Integrity, a consultancy firm in San Francisco, California.

Being unemployed and independent gives Bik the freedom to speak her mind, she says. "This one looks like nobody gave a fork about putting together a good science paper," she tweeted in March, with an accompanying figure panel that contained multiple duplicated

ADVANCED SUPER-SPOTTER TEST

Elisabeth Bik identified eight duplicate areas in these western blot images from separate experiments in four figures in a paper. We have simplified the labels. Can you spot the duplications? (Answers overleaf).



images. Last July, Bik commented on an image: "For those of you who did not get an NIH R01 grant around 2005, this is where that money was spent on instead."

But there is also risk, especially for someone who refers to herself as "blunt and snarky" on her Twitter biography. "At some point, I am afraid people will sue me," she says. She tries to keep her critiques to research papers, rather than accusing their authors. Bik has not faced a lawsuit, but has been harassed and has sometimes taken time off Twitter. One person e-mailed her former colleagues at Stanford arguing that she had abused her research grant funding by pursuing image integrity investi-

"We cannot, unfortunately, clone Elisabeth."

gations during work hours. (Bik says this was untrue.) Another posted personal information on PubPeer (now removed). "I've been called a bitch a couple of times," she says. "It comes with the work I do."

Because she posts under her real name, Bik says she errs on the side of caution, sometimes deciding not to flag cases online, especially those with blurry or low-resolution images. On her own science-integrity blog, many entries begin with some version of the phrase: 'This post is not an accusation of misconduct'. Suspicious images don't always point to corrupt actions, she says: researchers might have mistakenly uploaded a file twice when preparing figures, for instance. Then there are technical artefacts: membrane-thin slices cut sequentially from a piece of tissue can stick together along one edge and flip open butterfly-style, creating an apparent mirrored duplication.

Defects on an old microscope can create dark spots that seem the same on every image.

"She has a good track record," says Bernd Pulverer, chief editor of *The EMBO Journal*, who calls Bik a world leader in manual image screening. "The things she calls out are usually real issues."

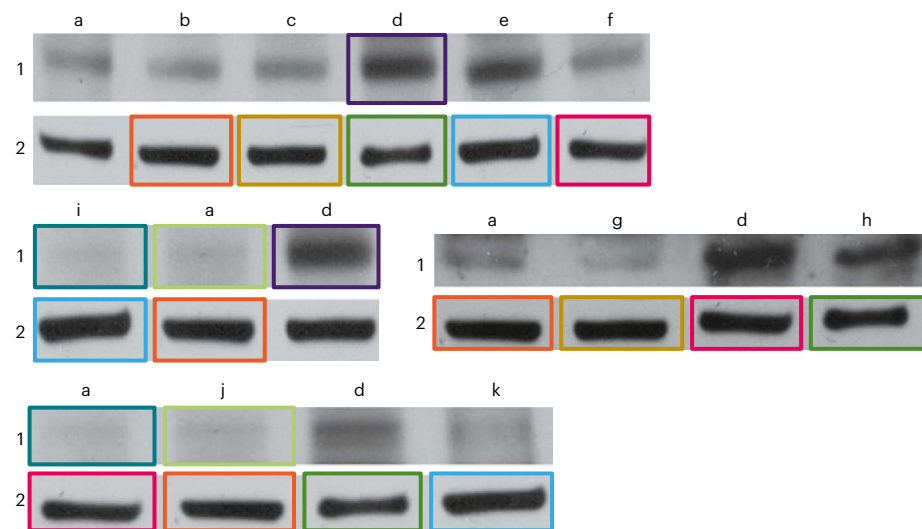
Public or private

Although many praise Bik for her work, some say the concerns shouldn't be aired in public before they are flagged privately to journals or research institutions. "It's very problematic," says Lauran Qualkenbush, president of the US Association of Research Integrity Officers. She says that, in cases in which foul play is suspected, a public outing might hinder investigative procedures by universities. "If someone did conduct research misconduct intentionally, and then they're alerted to the concern, it's a great opportunity for them to destroy evidence," she says.

Bik – in common with other image sleuths – says she's tried informing journals privately, but the case often seems to go nowhere or take too long to resolve. (She also notes that researchers have opportunities to destroy evidence even if investigations occur in private.) Between 2014 and 2015, Bik reported all 782 questionable papers from her 2016 *mBio* study directly to journals through e-mail. Some journals were unprepared for the sheer volume of Bik's reports. She flagged 348 papers of concern to *PLoS ONE* in a raft of 30 e-mails, each with 10 or 20 attachments. "That obviously created a backlog because we were not equipped to deal with it," says editor-in-chief Joerg Heber. Eventually, in 2018, the journal formed a three-person team dedicated to investigating image integrity and other publication ethics cases full-time. "We published around 100 retractions last year.

DID YOU SPOT THEM?

In the original paper, the figures were far apart. Finding duplicates between papers is even harder.



Many of these were among cases that had been raised by her," says Heber. The team is still working through Bik's original tips, as well as other cases. Bik gives *PLoS ONE* credit for its efforts, and says she receives regular notifications of *PLoS ONE* retractions and corrections that have stemmed from her leads. But with many of the nearly 800 cases in the *mBio* data set still unresolved, Bik's patience is wearing thin. "I can tell you that 60–70% have not been addressed after five years, so now, yes, I'm going to take it more publicly," she says.

The due-diligence process to check concerns with papers often takes much longer than people expect, Pulverer says. He and Heber note that waiting for responses and raw data from authors, and sometimes research institutions, can be time-consuming.

Bik says she realizes that investigations take time. But she argues that journals could use expressions of concern more quickly and frequently to notify other researchers of potential problems, while possibly years-long investigations are pending. Heber says *PLoS ONE* uses expressions of concern when it has gathered enough information to be concerned, but might hold off if an investigation is running smoothly, in favour of reaching a resolution such as a correction or retraction. *Nature's* editor-in-chief, Magdalena Skipper, says that expressions of concern, which alert readers to "serious concerns" with a paper, are "a formal and permanent part of the scientific record; as such, we endeavour to use them judiciously, adding them to papers once we have evidence that it is appropriate to do so".

These days, Bik typically reports her discoveries directly on PubPeer. Some journals and publishers track activity on the site, so she can reach journal editors and the public. "It's more important to flag these papers and not worry about what happens behind the scenes with these institutes," she says.

Many – including Bik – argue that combating image manipulation and duplication requires system-wide changes in science publishing, such as greater pre-screening of accepted manuscripts. "My preference is not to have to clean up the published literature, but to do it beforehand," says Rossner. He helped to introduce universal image pre-screening of accepted manuscripts at the *Journal of Cell Biology* nearly 20 years ago. At the EMBO Press, says Pulverer, journals have pre-screened accepted papers for faulty images since 2013. But most journals still do not pre-screen or (as with *Nature*) spot-check only a subset of papers before publication. "Image screening is not common right now," says Chris Graf, director of research integrity at the publisher Wiley.

But the tide is slowly turning; Wiley publishes a few journals that screen images, and is "preparing to launch a screening service" with the *Journal of Cellular Biochemistry* and the *Journal of Cellular Physiology*, Graf says. The journal *Science* has editorial coordinators who check accepted manuscripts for signs of image manipulation, but they don't have capacity to check for some issues, such as whether figures have been flipped, rotated or duplicated, says executive editor Monica Bradford.

A job for AI?

Many researchers say automation is the key to improving image integrity at a large scale. "We cannot, unfortunately, clone Elisabeth," says Daniel Acuna, a computer scientist at Syracuse University in New York, whose group is one of a handful working on algorithms to detect problematic images. Although Bik excels at finding duplicated images in a single paper, computers could help to find more duplications between papers by comparing hundreds of thousands or millions of papers – an unfeasible task for humans, he says. In 2018,

Acuna's team published on the bioRxiv preprint server preliminary results of an analysis that extracted 2 million images from 760,000 papers (D. E. Acuna *et al.* Preprint at bioRxiv <http://doi.org/dtp2>; 2018). It proved too computationally intensive to compare every image with every other, but the team looked at image reuse within and across papers by the same authors. After manually examining a sample of more than 3,700 of the matching images that the software flagged, the researchers identified 40 cases that they all agreed were probably fraudulent; almost half of these involved the same image being used to represent different results in different papers.

Current technology is good at detecting outright duplications, and flipped or rotated copies, says Bucci. His company, Resis, uses proprietary software to scan scientific manuscripts for its clients, which include journals and research institutions. But complex problems are tougher, such as two images that share a small overlapping area, but are otherwise completely different. Advances in machine learning could be the key to detecting these and other subtle patterns automatically, he says.

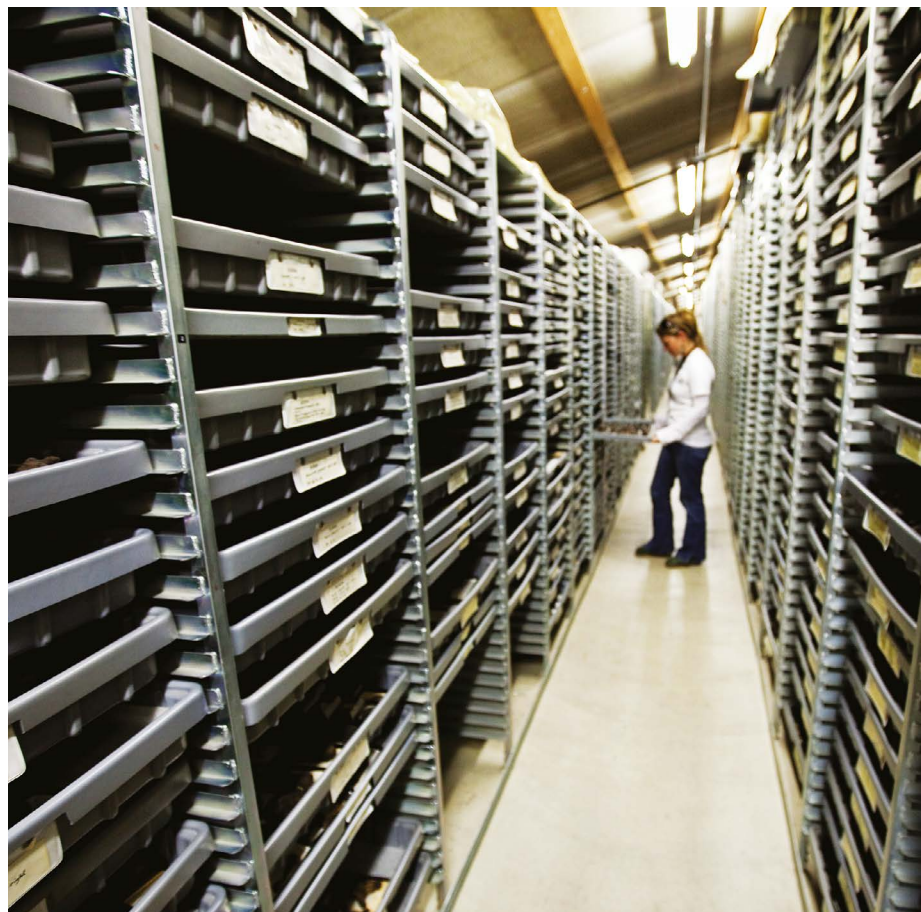
But better software will need more data. Machine-learning algorithms require training with an abundance of images that are known to contain duplications. Bik has shared with Acuna images from hundreds of 'dirty' and 'clean' papers from her 2016 study. And at the Humboldt University of Berlin, researchers funded by the publisher Elsevier are developing a searchable database of images from retracted papers. For now, the collection has fewer than 500 entries, and most are in the life sciences and medicine and contributed by Elsevier, so the team wants more publishers to participate. The publisher says that some of its journals are piloting image-checking software, and its goal is to provide all its journals with automated systematic checking.

Until recently, Bik was unimpressed by the software available. Now, she says, "I have full confidence that in the next two years, computers will be usable as a mass way of screening manuscripts." But both Bik and Acuna say that people will always need to check the results of such programs, especially to weed out instances where images can and should look similar in certain parts.

For now, Bik has plenty of work to do. This morning's tip from Belgium looks like it might be a hit. Some of the western-blot bands – normally fuzzy and rounded like tiny black caterpillars – sport unusually sharp, pixelated edges, she says; these could be an innocent artefact introduced when a picture is compressed to a smaller size, or could suggest the application of photo-editing tools. "I'm going to ask him for the rest of the paper," says Bik.

Helen Shen is a science journalist based in Sunnyvale, California.

Comment



A staff member views the fossil collection at La Brea Tar Pits museum in Los Angeles, California.

Store and share ancient rocks

Noah Planavsky, Ashleigh Hood, Lidya Tarhan, Shuzhong Shen and Kirk Johnson

Geological samples must be archived for all if we are to solve the riddles of Earth's complex history.

Geologists think they know the basics of Earth's history. Liquid water has flowed on the planet for 4 billion years¹. Tiny amounts of oxygen first gathered in the atmosphere about 2.3 billion years ago². And the planet went through many periods of climatic upheaval, from freezing completely 700 million years ago³ to warming so rapidly about 250 million years ago that more than 80% of marine species were lost^{4,5}. It has had many more ups and downs.

This story can be reconstructed using data wrestled from ancient rocks. But as geologists

learn more, our planet's tale is getting muddier rather than clearer. Controversies have erupted in the past two decades over many aspects of the chemical record of the early Earth, including the evolution of life, environments and past long-term climate (see 'Contentious timeline').

For example, variations in carbon-isotope ratios in carbonate rocks have conventionally been interpreted as recording drastic global environmental changes, including huge episodes of volcanism or bursts of oxygen⁶. By contrast, some researchers suggest that these same records have been changed over time by local environmental processes, and that they do not provide information about Earth's ancient history⁷. This debate can be resolved only by applying a variety of geological and chemical tools^{8,9} to the same samples used to generate the carbon-isotope results.

Attempts over the past decade to answer questions using better tools and larger databases have only amplified disputes. To make matters worse, too often, rock samples are not archived or shared. It is common for samples to be held by researchers in private collections instead of in accessible, curated institutional archives or museums. That's a problem, because different geoscience teams cannot check each other's work to test whether published results are robust and can be replicated.

We call on researchers, museums, funders, scientific societies and journals to ensure that all samples of sediment and sedimentary rock from which geochemical data have been produced and published are curated, archived and made available to members of the research community.

Reproducibility crisis

Geological records are complicated and hard to interpret. It is easy to reach contradictory conclusions, most commonly for the following four reasons.

Proxies and archives. Several geochemical methods can be used to infer past conditions such as temperature. The same method applied to different sedimentary rock types can lead to inconsistencies. For example, the ratio of heavy to light oxygen isotopes in chemical precipitates (such as chert, carbonate or apatite) tracks the seawater temperatures under which these minerals formed. But even in the same piece of rock, the reconstructed temperatures can be different depending on whether they are measured in a fossil or in a bulk aggregate of the entire rock sample. This is because rocks are inherently combinations of different minerals, which might have

Comment

formed during different stages of a rock's long geological history. The consequences for understanding past climates can be dramatic. For example, it is still not clear whether an interval of extreme heat killed marine organisms during the 'Great Dying' 250 million years ago. Sulfide toxicity, ocean acidification and carbon dioxide poisoning have also been proposed as possible mechanisms for killing off organisms at this time⁴.

Similarly, the question of whether oxygen levels were low enough to have delayed the emergence of animals for around 4 billion years – or most of Earth's history, thus addressing Charles Darwin's dilemma of why complex life appeared so late in the fossil record – depends on which rocks are studied and what analytical methods are used⁸. For example, an analysis of gas bubbles in sedimentary rocks⁹ has suggested that atmospheric oxygen levels on Earth's surface would have been high enough to support animals as early as 2.6 billion years ago. However, this clashes with a compelling body of evidence indicating that atmospheric oxygen concentrations were vanishingly low at this time^{10,11}. Refining such proxies is extremely challenging when different teams cannot work on the same samples.

Geographical and temporal variation. Rock samples that are used to tackle the same research question are often collected from different places, where the rocks were deposited at various times and in vastly different environments. This can result in completely distinct answers. For example, mercury enrichments in sediments are used as a tracer of large episodes of volcanic activity and their links to mass extinction events¹². However, mercury enrichments can also result from wildfires or from local depositional conditions that lead to heavy-metal uptake by sedimentary organic matter¹². Furthermore, diverse geographical settings can record mercury enrichments differently, depending on aspects such as water depth, dissolved oxygen concentrations, the rate of sediment deposition and the type and location of the volcanoes themselves^{12,13}. All of this can lead to spurious correlations between volcanism and extinction events. It is difficult to disentangle signals of global changes in the Earth system through time from local environmental variability using only reported geochemical data sets.

Analytical reproducibility. Experiments can be hard to repeat even if rocks are pristinely preserved. Measurements are routinely checked against those of geochemical standard materials, the compositions of which are internationally validated. Yet there is always the possibility of errors during analysis. These can arise from differences in sample preparation (such as in rock-crushing techniques or in the type of acid used to prepare a sample)

and instrumentation (machine type, tuning) to variations in laboratory conditions. For instance, boron-isotope measurements on marine carbonates are one of the key tools used to reconstruct atmospheric CO₂ levels¹⁴. Various approaches to making such measurements can lead to CO₂ estimates that differ by more than 400 parts per million^{14,15} – roughly equivalent to the total concentration of CO₂ found in the atmosphere today.

Contamination and alteration. As sediments become rocks, they undergo many processes that can alter the geochemical signals of where and how they formed. Sediments laid down on sea floors or lake bottoms can experience changes in water level or salinity, for example if they are flushed with meltwater. Hydrothermal processes and heat at depth might leach chemicals from the rock and alter the mineral composition.

Rocks collected near the surface can be altered by groundwater or other contaminants, such as oil used to drill cores. For example, organic remains in rocks once thought to be evidence for oxygen production by pioneering photosynthetic microorganisms 2.7 billion years ago are now acknowledged to be probable contamination from the modern petroleum products used to drill the rocks from the ground¹⁶. Similarly, debate is raging over whether the chemical composition of ancient rocks records microbial oxygen production extending as far back as 3 billion years ago, or whether those rocks have been compromised by contact with recent groundwater¹⁷.

Precious prizes

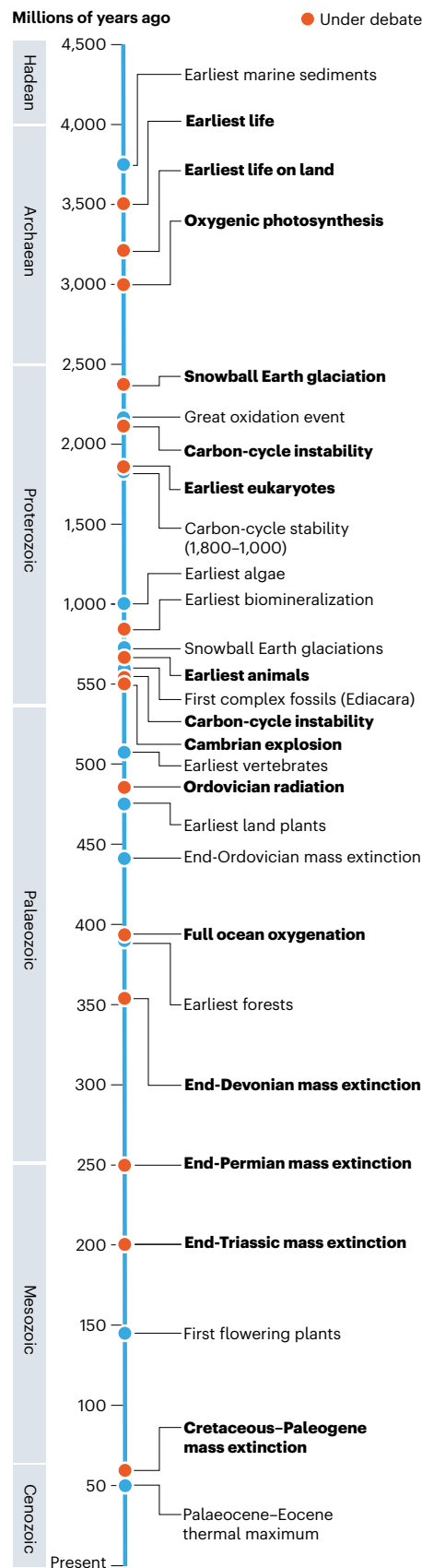
Without the ability to access and remeasure samples, it can be challenging to work out whether disparities in results and views stem from complexity in Earth's history, from sampling of rocks with different levels of alteration or from analytical issues. Yet sample archiving is not part of the standard protocol for inorganic or organic geochemical work, nor for some palaeoclimate work (other than, for example, ocean drill-core or ice-core samples, which are stored).

Why has this situation arisen? Many scientists are reluctant to share samples they have struggled hard to collect. After all, there are high costs associated with fieldwork on outcrops and drilling programmes. Research groups might want to perform multiple geochemical studies on a single set of samples, and this takes time. Large geochemical studies that use unconventional isotope systems can take several years to extract a data set¹⁸.

Other obstacles to archiving samples include how to fund archiving, where to store samples and how they are to be managed. Clearly, no single museum can hold all geological and geochemical samples. Museums would need to increase staff, space and funds

CONTENTIOUS TIMELINE

Earth's environmental history has been reconstructed with data wrestled from ancient rocks. Some events remain hotly debated. Archiving and sharing of rock samples enable published work to be tested and replicated.



SOURCE: N. PLANAVSKY ET AL.

for such collections.

Initiatives for archiving materials in other fields could serve as models. These include the Global Genome Initiative, a shared data protocol for frozen tissue repositories (see go.nature.com/3f4erur), and the Integrated Digitized Biocollections project for biological digital data (www.idigbio.org). Global databases of these sample archives and their accessory information, building on initiatives such as the International Geo Sample Number (IGSN; www.igsn.org), will also be needed to assign unique identifiers and maintain inter-collection records.

Some Earth-science fields already deposit samples in publicly accessible museums. For instance, palaeontologists have been required to do so for samples formally described in scientific publications for more than 150 years. Likewise, museums hold type specimens of fossils, meteorites and biological samples. Well-funded drilling projects also have strict archiving policies and well-curated core libraries, such as that for the International Ocean Discovery Program (see go.nature.com/2xoumhh).

The FAIR data initiative offers strict guidelines on data archiving and has been adopted by many journals that publish Earth- and environmental-science research, including *Science* and *Nature* (see go.nature.com/2wv2jxd). Although the recommended best practices of this initiative already include sample archiving, this is not yet strictly implemented as a formal requirement for publication.

Forging consensus

Together, researchers, natural history museums, journal editors, scientific societies and funding agencies must develop and implement standardized archival policies. We recommend that the following steps are taken.

Geochemical researchers should routinely send their samples to museums. To encourage buy-in, we suggest an embargo period for delaying new studies by other research groups on each set of samples from which geochemical data have been published. Geochemists must also work with museums to broaden the conventional definition of collections to include a range of different materials, from fist-sized specimens to rock fragments, powders and mineral grains. Geochemists should work with custodians of protected lands to encourage the inclusion of archival policies and procedures for geochemical samples collected under research permits.

Natural history museums should broaden their mission to archive and curate geological samples. They should assign unique identifiers that can be logged in digital databases. Curators must decide how much of a sample can be withdrawn, because geochemical tests are destructive. Where resources are tight, museums will need to evaluate the

spatial, financial and scientific capacity of collections, and determine which samples are most essential to curate.

Scientific societies must tackle the question of what constitutes an acceptable repository. For instance, the Meteoritical Society's Committee on Meteorite Nomenclature does this. Scientific societies such as the Geochemical Society in Washington DC and the European Association of Geochemistry in Aubière, France, should begin to recommend suitable institutions.

Recent decades have demonstrated that rapid changes in data archiving are possible when clear guidelines – and editorial mandates – are in place. So we would like to see journals go further in supporting the FAIR data initiative, by making requests to archive samples and assignment of database unique identifiers mandatory for publication.

Many scientific journals regulate data archiving using a checklist. We recommend that this practice be implemented for sample archiving, and that repository-issued sample identifiers (as well as unique identifiers assigned by inter-institutional database efforts such as the IGSN) be included in each paper. All major changes to a field take time to develop, and changes at the editorial level

“We estimate that roughly 200,000 new sedimentary geochemical samples are analysed each year.”

can help to nudge them along. Journals could implement these policies on a relatively short timescale, as long as exceptions are initially made to the archiving mandate when requests for sample deposition are declined.

Funding agencies should require that researchers' grant proposals include sample archival procedures and that budgets include curation fees. Critics might argue that archiving will decrease the money available for other scientific endeavours. In our view, a sample stewardship plan should be viewed as equivalent to budget-line items for data archiving, publishing fees or institutional overhead costs that support other essential components of the research workflow.

We strongly recommend against setting universal fees. Samples will vary widely in nature and size, from kilogram-scale samples to micrograms of separated minerals. So the cost to museums will likewise depend on institutional resources and expertise. However, we have confidence that museums, working with funding agencies and researchers, will ensure that fees are self-regulating.

Collections of palaeontological samples provide an analogue for the practices needed. They also show that large-scale archiving is possible.

The Invertebrate Paleontology Division of the Yale Peabody Museum of Natural History in New Haven, Connecticut, for instance, holds about 4.5 million specimens and takes in more than 2,000 samples a year, on average. As well as its curatorial researchers, the division is supported by two full-time staff members, one of whom handles the new acquisitions.

We estimate that roughly 200,000 new sedimentary geochemical samples are analysed each year. We therefore reiterate that curation fees – even modest ones – should be incorporated into the budgets of research-grant proposals. Regardless of the current availability of space and curatorial support in individual museums, extra funds will be needed to meet the demand for archiving sedimentary geochemical samples.

The guidelines we offer will need to be discussed and revised by the community and institutions. Nonetheless, all best practices must rest on a shared commitment – to ensure that scientific data are not divorced from scientific samples.

The authors

Noah Planavsky is an associate professor in geochemistry at Yale University, New Haven, Connecticut, USA, and assistant curator of mineralogy and meteoritics at the Yale Peabody Museum of Natural History. **Ashleigh Hood** is a lecturer in sedimentology at the University of Melbourne, Parkville, Australia. **Lidya Tarhan** is an assistant professor in palaeontology and sedimentology at Yale University, New Haven, Connecticut, USA. **Shuzhong Shen** is a professor in palaeontology and stratigraphy at Nanjing University, Nanjing, China. **Kirk Johnson** is the director of the Smithsonian's National Museum of Natural History, Washington DC, USA. e-mail: noah.planavsky@yale.edu

1. Cavosie, A. J., Valley, J. W. & Wilde, S. A. *Earth Planet. Sci. Lett.* **235**, 663–681 (2005).
2. Farquhar, J., Bao, H. & Thiemens, M. *Science* **289**, 756–758 (2000).
3. Hoffman, P. F. et al. *Sci. Adv.* **3**, e1600983 (2017).
4. Fan, J.-X. et al. *Science* **367**, 272–277 (2020).
5. Sun, Y. D. et al. *Science* **338**, 366–370 (2012).
6. Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J. & Chivas, A. R. *Nature* **537**, 535–538 (2016).
7. Allwood, A. C., Rosing, M. T., Flannery, D. T., Hurowitz, J. A. & Heirweh, C. M. *Nature* **563**, 241–244 (2018).
8. Cole, D. B. et al. *Geobiology* **18**, 260–281 (2020).
9. Steadman, J. A. et al. *Precamb. Res.* **340**, 105722 (2020).
10. Luo, G. et al. *Sci. Adv.* **2**, e1600134 (2016).
11. Lyons, T. W., Reinhard, C. T. & Planavsky, N. *J. Nature* **506**, 307–315 (2014).
12. Grasby, S. E., Them, T. R., Chen, Z., Yin, R. & Ardakani, O. H. *Earth-Sci. Rev.* **196**, 102880 (2019).
13. Percival, L. M. E. et al. *Am. J. Sci.* **318**, 799–860 (2018).
14. Foster, G. L. et al. *Chem. Geol.* **358**, 1–14 (2013).
15. Hennehan, M. J. et al. *Proc. Natl. Acad. Sci. USA* **116**, 22500–22504 (2019).
16. French, K. L. et al. *Proc. Natl. Acad. Sci. USA* **112**, 5915–5920 (2015).
17. Albut, G. et al. *Geochim. Cosmochim. Acta* **265**, 330–353 (2019).
18. Isson, T. T. et al. *Geobiology* **16**, 341–352 (2018).

News & views

Astronomy

A glimpse inside δ Scuti stars

József M. Benkő & Margit Paparó

Patterns in the vibrations of stars produce a sort of natural music that offers clues to the stars' internal structure. Astronomers have identified such patterns for some δ Scuti stars, a group for which this music had been elusive. **See p.147**

Our knowledge of the stars is based almost exclusively on the study of their light. But the light that reaches us originates from their upper layers – we can't see inside. There is, however, a tool we can use to look into the interior of the stars: asteroseismology. On page 147, Bedding *et al.*¹ report that a subgroup of the enigmatic δ Scuti stars exhibits regular pulsations that will finally enable the stars to be probed using this tool.

Inside a star, gravity and gas pressure compete with each other. If the two are in balance, the star is in equilibrium, but if one increases more than the other, the star contracts or expands. Hot gas spheres such as stars can show characteristic periodic oscillations in which the star pulsates in this way. These characteristic oscillations, called eigenmodes, are standing waves, like the standing sound waves responsible for the sounds of musical instruments such as violins and oboes. The eigenmodes are determined by the physics of the oscillating system.

Asteroseismology is the study of these stellar oscillations². The idea is similar to that of seismology, in which the interior structure of our planet is inferred from earthquakes. Each star can have different – and often very large numbers of – eigenmodes, depending on its internal structure. Oscillations that have different periods or frequencies are sensitive to physical conditions in different regions inside the stars. The more eigenmodes that can be determined from observations, the more detailed will be our map of the internal structure.

Oscillations produce brighter and darker areas (corresponding to higher and lower pressures and temperatures; Fig. 1) on the star's surface². However, we cannot resolve the surfaces of stars, apart from that of our

Sun and a few other special cases (see ref. 3 and references therein, for example). Only their total brightness can be measured. The complicated distribution and variation of surface brightness results in an equally complex temporal variation in the total brightness. By measuring the brightness of a star, a photometric time series known as the light curve is obtained.

For asteroseismology, then, we need the following steps. To obtain the surface-brightness distribution from the measured light curves, the frequency of the brightness

variations must be determined. Next, we must work out how these frequencies correspond to the eigenmodes expected from theoretical models, a process called mode identification. If the mode identification is successful, actual asteroseismology can begin by determining key physical parameters such as stellar mass and age. Then the ultimate goal of asteroseismology can follow: obtaining the total seismic inversion, which means the detailed determination of stratifications of the pressure, temperature and chemical composition inside the star.

For decades, researchers have made tremendous efforts to obtain valuable asteroseismic data sets. Extensive observation campaigns were carried out using ground-based telescopes, but inevitable variations in detectors and weather conditions affected the data. Space missions (such as CoRoT, Kepler and TESS) delivered the real breakthrough. Thanks to the missions' long, homogeneous and evenly sampled light curves, and the precision of the collected data, asteroseismology has now been successfully applied to thousands of stars across several stellar types that have different internal structures^{4–6}.

But the abundant star type known as δ Scuti (ref. 7), named after a star in the constellation Scutum, has remained one of the exceptions. Stars of this type have a slightly larger mass

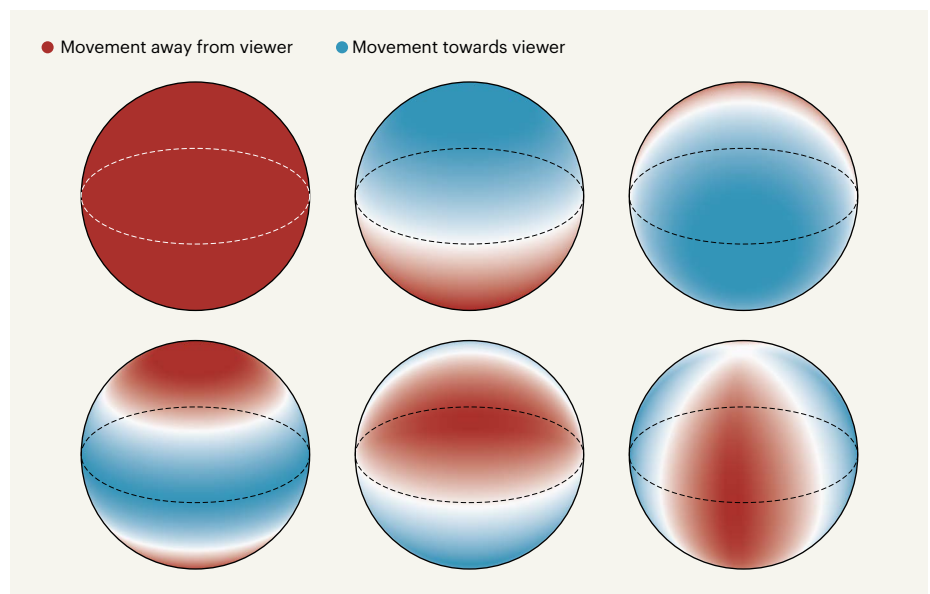


Figure 1 | Simple modes of stellar oscillations. Pressure oscillations in stars occur in many different combinations of characteristic patterns and frequencies, such as the simple examples shown here. These oscillations change the brightness of stars, and offer clues about the physical conditions inside them. The oscillatory modes of an important family of stars known as δ Scuti stars have been difficult to identify. Bedding *et al.*¹ have identified a group of δ Scuti stars that pulsate at a high rate and with regular patterns of frequency that agree with theoretical predictions. This allowed the oscillation modes to be identified.

than that of our Sun, but their inner structure is very different. They have long been known to have complicated, low-amplitude light curves⁷. One might think that the large number of detected frequencies would make these stars ideal targets for asteroseismology.

The theoretical models of δ Scuti stars predict many possible excited eigenmodes and corresponding frequencies. In fact, there are many more such frequencies in models than have been observed⁸, and usually we do not know which of the possible modes are seen. If there were some regular structure to the frequencies (such as frequencies with comb-like regular differences), we would have a better chance of identifying them. But the theoretical models generally do not predict regular frequency structure for these stars.

Bedding *et al.* have identified a special subgroup of δ Scuti stars that pulsate at higher frequencies than do most such stars. For this subgroup, both theory and observations suggest the existence of regular frequency structures. Other researchers have previously found regular structures in observed data for some δ Scuti stars (see refs 9–14, for example), but did not identify the oscillating modes conclusively, if at all. Bedding *et al.* provide unambiguous mode identification for a uniform and relatively large sample of these stars.

A key factor in the authors' success is that many of the stars in the subgroup rotate more slowly than do other δ Scuti stars. (Alternatively, it could be that some of the stars are observed almost pole-on, resulting in apparently small rotation velocities.) Theoretical models predict that the frequency spectra of stars that have low rotation velocities are less complicated than those with higher rotation speeds¹⁴, which makes it easier to recognize their regular frequency structures. Bedding *et al.* not only identified these structures, but also associated the frequencies with the corresponding eigenmodes.

Sky surveys now and in the near future will target many thousands of δ Scuti stars, including many that are similar to those described by Bedding and co-workers. This is not merely an opportunity to understand the physics of a special group of δ Scuti stars better. The authors show that these are young stars, which means that they can be used as tracers to estimate the age of open star clusters or of young stellar associations in our Galaxy. In this way, we might learn more about the evolution of the Milky Way. Bedding and colleagues' study is therefore not the last word on δ Scuti stars. Rather, it opens up avenues of investigation for this important stellar group.

József M. Benkő and Margit Paparó

are at Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, H-1121 Budapest, Hungary.
e-mail: benko@konkoly.hu

1. Bedding, T. R. *et al.* *Nature* **581**, 147–151 (2020).
2. Aerts, C., Christensen-Dalsgaard, J. & Kurtz, D. W. *Asteroseismology* (Springer, 2010).
3. Paladini, C. *et al.* *Nature* **553**, 310–312 (2018).
4. García, R. A. & Ballot, J. *Living Rev. Sol. Phys.* **16**, 4 (2019).
5. Hekker, S. & Christensen-Dalsgaard, J. *Astron. Astrophys. Rev.* **25**, 1 (2017).
6. Corsico, A. H., Althaus, L. G., Miller Bertolani, M. M. & Kepler, S. O. *Astron. Astrophys. Rev.* **27**, 7 (2019).
7. Breger, M. *ASP Conf. Ser.* **210**, 3–42 (2000).
8. Breger, M., Lenz, P. & Pamyatnykh, A. A. *Mon. Not. R. Astron. Soc.* **396**, 291–298 (2009).
9. Zwitter, K. *et al.* *Astron. Astrophys.* **533**, A133 (2011).
10. Breger, M. *et al.* *Mon. Not. R. Astron. Soc.* **414**, 1721–1731 (2011).
11. Zwitter, K. *et al.* *Astron. Astrophys.* **552**, A68 (2013).
12. García Hernández, A. *et al.* *Astron. Astrophys.* **559**, A63 (2013).
13. Paparó, M., Benkő, J. M., Hareter, M. & Guzik, J. A. *Astrophys. J. Suppl. Ser.* **224**, 41 (2016).
14. Unno, W., Osaki, Y., Ando, H., Saio, H. & Shibahashi, H. *Nonradial Oscillations of Stars* (Tokyo Univ. Press, 1989).

Immunology

Brain–spleen connection aids antibody production

Flurin Cathomas & Scott J. Russo

Elucidating how the brain controls peripheral organs in the fight against infection is crucial for our understanding of brain–body interactions. A study in mice reveals one such pathway worthy of further investigation. **See p.204**

Interactions between the mind and the body have sparked the interest of scientists and philosophers for centuries. In ancient Greece, the physician Galen described the spleen as being the source of black bile, which was thought to cause melancholy when secreted in excess. Today, research is uncovering complex ways in which the brain and body interact to affect diverse aspects of health, from mood to immune function. The spleen aids immune defences by functioning as part of the lymphatic system; the organ is a major hub of activities needed to initiate responses in the adaptive branch of the immune system, which handles defences that are tailored to a specific disease-causing agent.

The spleen is a target of top-down control from the brain¹. Zhang *et al.*² have taken our understanding of brain–spleen connections to the next level by revealing on page 204 an aspect of top-down control that regulates the adaptive immune system.

The spleen's contribution to immune responses occurs mainly in its white-pulp region, where immune cells that have arrived from elsewhere in the body present peptide fragments called antigens to immune cells called T cells. If a T cell binds to and recognizes such an antigen, which might indicate the presence of an abnormal cell or a foreign invader, this activates the T cell, which in turn activates immune cells called B cells. B cells differentiate to form plasma cells (Fig. 1) that secrete antibodies specific for the antigen presented, and these antibodies are released into the bloodstream to fight infection³.

Spleen activity is controlled by the autonomic nervous system – a part of the nervous

system that regulates organs. More specifically, the spleen is controlled mainly by the sympathetic branch of the autonomic nervous system, which is associated with the 'fight-or-flight' response⁴. However, little was known previously about possible upstream brain regions that might connect to the autonomic nervous system in the spleen to control it and, by extension, adaptive immunity. An earlier study in mice⁵ revealed that stimulation of a brain region called the ventral tegmental area, a part of the brain's reward circuit, boosts immune responses and protection against harmful bacteria.

Zhang and colleagues developed a surgical technique to remove nerves from the spleen in mice. This mainly removed inputs from the autonomic nervous system and prevented top-down control from the brain to the spleen. After surgery, the animals were injected with an antigen. Plasma cells that made antibodies targeting that antigen arose in abundance in control mice that had undergone a 'sham' operation that did not remove nerves. Such an increase did not occur in the denervated mice, indicating that splenic-nerve activity regulates the formation of plasma cells and thus adaptive immunity.

The authors investigated which molecular mechanisms might be needed for plasma-cell formation in this context. They studied the expression of various types of receptor that can bind the neurotransmitter molecule acetylcholine, which is a key signalling component of the autonomic nervous system. Zhang *et al.* report that B cells express a type of acetylcholine receptor called a nicotinic receptor, and the authors pinpointed protein subunits

of this receptor, including one called *Chrna9*. To test the role of nicotinic receptors containing *Chrna9* in plasma-cell formation, Zhang *et al.* transplanted haematopoietic stem cells, which can generate immune cells, into mice that had undergone a treatment to remove their own haematopoietic stem cells. When the transplanted stem cells came from mice engineered to lack the gene encoding *Chrna9*, these animals generated fewer plasma cells after an injection of antigens than did animals that received antigen injections and transplants of stem cells with the gene intact. This result indicates that plasma-cell formation requires the presence of nicotinic receptors.

When a type of T cell called a CD4⁺ T cell is activated by antigen recognition, it secretes acetylcholine in response to the hormone noradrenaline⁶. The authors reveal that such T cells serve as a ‘relay’ between the release of noradrenaline from the splenic nerve and the subsequent acetylcholine-dependent⁶ formation of plasma cells (Fig. 1).

To map the neural circuit that connects the spleen and brain, the authors used a method termed retrograde tracing, which relies on monitoring the expression of a fluorescent protein encoded by a virus that can ‘jump’ across the synapses that connect neurons. This enabled Zhang and colleagues to track all upstream inputs to a given nerve cell in the spleen. The authors thereby identified two key brain regions (the central nucleus of the amygdala and the paraventricular nucleus of the hypothalamus) that contain neurons that connect to splenic nerves. These regions are major centres involved in the response to psychological stressors such as fear or threatening situations⁷, and they have essential roles in regulating the production of neuroendocrine hormones, for example, by a pathway called the hypothalamic-pituitary-adrenal axis⁸.

One population of nerve cells in these two regions releases the hormone corticotropin, which is thought to have a key role in initiating the body’s response to stress⁹. To determine whether corticotropin-producing neurons affect the spleen, Zhang *et al.* stimulated these neurons using a technique called optogenetics, and assessed whether this affected the activation of splenic nerves by monitoring their firing using electrophysiological recording. This provided crucial functional evidence for a brain–spleen connection, because such stimulation increased the firing of splenic-nerve cells. The authors also report that the inhibition or ablation of corticotropin-producing neurons in either of the two brain regions impaired the formation of plasma cells after antigen injection. Conversely, activation of the neurons stimulated such plasma-cell formation.

Although these circuit-based experimental approaches provide key proof for the existence of the brain–spleen axis, the authors also needed to test their model using suitable interventions

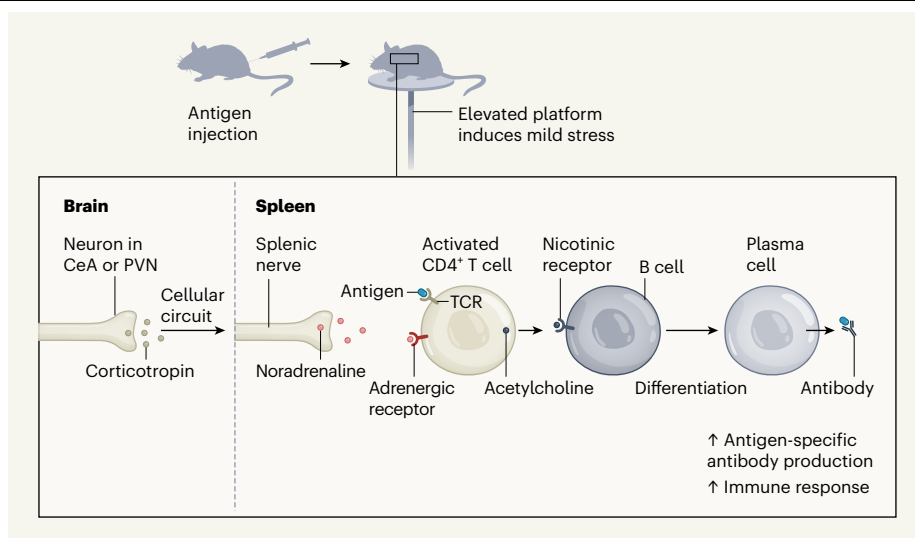


Figure 1 | Brain control of antibody production. Zhang *et al.*² describe a circuit between the brain and the spleen that aids immune defences. The authors injected animals with an antigen (a peptide fragment) that can be recognized by immune cells. Placing the animal on a high platform activated neurons that produce the molecule corticotropin. These neurons are located in brain regions that respond to stress, called the central amygdala (CeA) and the paraventricular nucleus (PVN) of the hypothalamus. A cellular circuit connects these activated neurons to the splenic nerve and drives it to release the molecule noradrenaline. An immune cell termed a CD4⁺ T cell is activated when its T-cell receptor (TCR) binds to antigen. When such a cell encounters the noradrenaline released in the spleen (which binds to what is termed an adrenergic receptor), this leads the T cell to secrete the molecule acetylcholine⁶. This molecule binds to a nicotinic receptor on an immune cell called a B cell, causing it to differentiate into a plasma cell. The plasma cell boosts immune defences by making antibodies that recognize the specific antigen that activated the T cell.

that activate the ‘stress centres’ in the brain. However, neurons in the central nucleus of the amygdala and the paraventricular nucleus function in a pathway that causes the adrenal gland to secrete the hormone glucocorticoid in response to stress, and glucocorticoids are potentially immunosuppressive¹⁰.

The authors therefore considered whether the concentration of glucocorticoids secreted by the adrenal gland might depend on the severity of the stress. To avoid possible glucocorticoid-driven immunosuppression that might interfere with their analysis of antibody production, Zhang *et al.* studied mice that had been placed on an elevated, transparent platform; this provided a behavioural situation that induced only moderate stress. Following antigen injection, this scenario, but not another set-up that caused more-severe stress, led to the generation of antigen-specific antibodies. The authors showed that this antibody production depends on corticotropin-producing neurons in the brain circuit that they had described.

There is growing evidence that dysregulation of the immune system has a bottom-up role in promoting several behaviours relevant to neuropsychiatric disorders¹¹. Zhang and colleagues’ study provides insights in the other direction – how the brain exerts top-down control of immune-system function. Future research will be needed to investigate whether this particular brain–spleen circuit exists in humans. The authors’ work opens up the exciting possibility that activating

certain brain regions (through behavioural interventions or by selective stimulation using neuromodulatory techniques such as transcranial magnetic stimulation) could modulate the immune system. To return to Galen, he was right that the spleen is a key site of connection between the brain and the body, but his ideas about how the spleen induces melancholy now give way to this new perspective on how the mind might modulate resilience-promoting antibodies.

Flurin Cathomas and **Scott J. Russo** are in the Nash Family Department of Neuroscience and the Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.
e-mails: flurin.cathomas@mssm.edu;
scott.russo@mssm.edu

1. Mebius, R. E. & Kraal, G. *Nature Rev. Immunol.* **5**, 606–616 (2005).
2. Zhang, X. *et al.* *Nature* **581**, 204–208 (2020).
3. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. *Nature Rev. Immunol.* **15**, 160–171 (2015).
4. Jung, W. C., Levesque, J. P. & Ruitenberg, M. J. *Semin. Cell Dev. Biol.* **61**, 60–70 (2017).
5. Ben-Shaanan, T. L. *et al.* *Nature Med.* **22**, 940–944 (2016).
6. Rosas-Ballina, M. *et al.* *Science* **334**, 98–101 (2011).
7. Davis, M. *Annu. Rev. Neurosci.* **15**, 353–375 (1992).
8. Smith, S. M. & Vale, W. W. *Dialogues Clin. Neurosci.* **8**, 383–395 (2006).
9. Peng, J. *et al.* *Front. Neuroanat.* **11**, 63 (2017).
10. Coutinho, A. E. & Chapman, K. E. *Mol. Cell. Endocrinol.* **335**, 2–13 (2011).
11. Cathomas, F., Murrough, J. W., Nestler, E. J., Han, M. & Russo, S. J. *Biol. Psychiat.* **86**, 410–420 (2019).

This article was published online on 29 April 2020.

Complex condensations get cells organized

Chiu Fan Lee

Liquid-like organelles in cells form when key constituents reach a certain concentration and then condense. Evidence now indicates that the concentration at which condensation occurs can vary, contrary to previous assumptions. **See p.209**

Water transitions from a liquid to a gas phase as it reaches its boiling point. Similarly, proteins in cells can transition from freely mixing in the cytoplasm or its nuclear equivalent, the nucleoplasm, to condensing into a concentrated liquid-drop phase once they reach a threshold concentration¹. This saturation concentration has been assumed to be an invariant quantity, but, on page 209, Riback *et al.*² demonstrate that this assumption is invalid. Much as the boiling point of water varies depending on pressure, the saturation concentration depends on the concentrations of the proteins involved.

Condensation of molecules into a liquid-like droplet – a process called phase separation – is a well-studied physical phenomenon, which can be caused by mutual attractions between proteins or other molecules. But many biological studies of phase separation so far have used simple model systems, rather than complex living cells. Riback and colleagues reasoned that the idea of a single fixed saturation concentration might have arisen because of

the use of simple systems.

In cells, phase separation produces liquid-like organelles called biomolecular condensates³. One such condensate is the nucleolus, in which the ribosome machinery involved in protein synthesis is made. Riback *et al.* set out to examine saturation concentration in cells by studying the protein

“The reputation of the nucleolus as the ribosome factory might be even more pertinent than people thought.”

nucleophosmin 1 (NPM1), which is a key driver of nucleolus formation^{4,5}. The group found that increasing the overall concentration of NPM1 in cells increased the corresponding saturation concentration at which the nucleolus forms in the nucleoplasm. Likewise, increasing the concentration of key proteins altered the

saturation concentration of stress granules – another type of liquid-like organelle.

Next, the authors showed that the variability of saturation concentration is caused by distinct interactions between a condensate's components. Rather than molecules of the same protein interacting during condensation, which might produce a fixed saturation concentration (NPM1 binding to other molecules of NPM1, for instance), the group found that phase separation depends on heterotypic interactions between different proteins in the condensate. As the concentrations of different proteins alter, the free energy of the nucleoplasmic mixture – the thermodynamic quantity that dictates how the components in the cell system are partitioned by phase separation – can change in a complicated manner, leading to changes in saturation concentration.

Biomolecular condensates are often intricately linked to cell functions⁶. Riback and colleagues went on to show how heterotypic interactions are exploited by nucleoli to facilitate the processing of ribosomal RNA, which makes up part of the ribosome. They found that phase-separating proteins such as NPM1 and another protein, SURF6, interact freely with immature forms of ribosomal RNA, but not as well as with more mature forms of the molecule. This leads to the mature RNA being expelled from the liquid-like nucleolus (Fig. 1). This finding highlights that nucleoli might not only act to concentrate key molecules and facilitate biochemical reactions, but also possess an underlying conveyor-belt mechanism to ensure a continuous and smooth production process. Hence, the reputation of the nucleolus as the ribosome factory might be even more pertinent than people thought⁷.

Riback and colleagues complemented each of their experimental findings theoretically, using methodology borrowed from equilibrium physics – the premise that there is no flow of energy into or out of a system. However, the environment of the cell interior, with its many processes driven by energy-carrying ATP molecules, is far from existing in equilibrium. As such, it is remarkable that the authors' close-to-equilibrium theory matches their real-world observations. I think that, although the picture laid out by Riback and colleagues is a valuable starting point, the reality will inevitably be more complex. Establishing a quantitative connection between experiments and theory will require further development of our theoretical understanding of non-equilibrium phase separation, which is still in its infancy^{8,9}. The fact that physicists do not know much about phase separation in non-equilibrium regimes should not be viewed as a drawback in the study of biomolecular condensates, however. Instead, it signposts a golden opportunity for life scientists, bioengineers and physicists to work closely together to expand our understanding of this complex phenomenon.

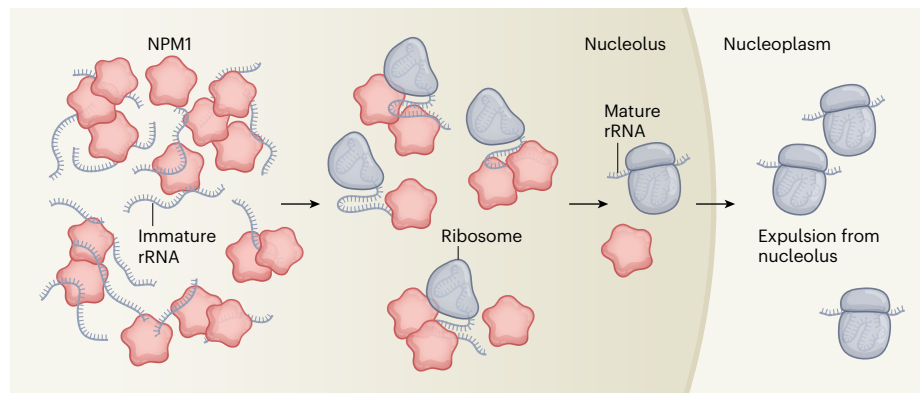


Figure 1 | Protein–RNA interactions control biological processes in the nucleolus. Riback *et al.*² report that complex interactions between different molecules govern the formation of liquid-like organelles such as the nucleolus, and can also regulate organelle function. The ribosome is a protein-synthesizing machine that is assembled from protein and RNA subunits in the nucleolus. The authors demonstrate that the proteins nucleophosmin 1 (NPM1) and SURF6 (not shown), which are key for formation of the nucleolus, interact freely with immature ribosomal RNA (rRNA). But as the rRNA becomes properly folded and incorporated into the ribosome, these interactions cease, and so the mature ribosomal RNA is expelled from the organelle into the surrounding nucleoplasm.

Chiu Fan Lee is in the Department of Bioengineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, UK. e-mail: c.lee@imperial.ac.uk

1. Israelachvili, J. N. *Intermolecular and Surface Forces* 3rd edn (Elsevier, 2010).
2. Riback, J. A. *et al. Nature* **581**, 209–214 (2020).
3. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. *Nature Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

4. Feric, M. *et al. Cell* **165**, 1686–1697 (2016).
5. Mitrea, D. M. *et al. eLife* **5**, e13571 (2016).
6. Shin, Y. & Brangwynne, C. P. *Science* **357**, eaaf4382 (2017).
7. Olson, M. O. J., Dundr, M. & Szebeni, A. *Trends Cell Biol.* **10**, 189–196 (2000).
8. Berry, J., Brangwynne, C. P. & Haataja, M. *Rep. Prog. Phys.* **81**, 046601 (2018).
9. Weber, C. A., Zwicker, D., Jülicher, F. & Lee, C. F. *Rep. Prog. Phys.* **82**, 064601 (2019).

This article was published online on 6 May 2020.

Atmospheric science

Airborne particles might grow fast in cities

Hugh Coe

Nanoscale particles have been observed to form and grow in the atmospheres of many cities, contradicting our understanding of particle-formation processes. Experiments now reveal a possible explanation for this mystery. **See p.184**

On page 184, Wang *et al.*¹ report observations of the rapid growth of newly formed atmospheric particles through the condensation of ammonium nitrate under conditions typical of many urban environments in wintertime. The observations were made in a chamber in a laboratory, but the authors convincingly argue that similar conditions can occur transiently in megacities. The findings fill a major gap in our knowledge of particle growth rates in cities.

Particulate matter is a key factor in the air quality of many of the world's megacities because it has been directly linked to multiple non-communicable diseases (see go.nature.com/2w49q1t). It also substantially affects regional climate through its interactions with solar radiation and clouds². Particle-formation processes are important in the air above large cities because they replenish the particle population, determine the total particle-number concentration and can act as 'seeds' for cloud formation. We therefore need to know how particles form and grow in order to predict the effects of particulate matter on health and regional climate.

Although our knowledge of particle formation has improved over the past few years^{3,4}, our understanding of the early stages of particle growth – particularly the crucial step in which an initial cluster of molecules grows large enough to become an actual particle – cannot explain why new particles form in megacity environments⁵. The persistence of newly formed clusters depends on the ratio of the condensation sink (the rate at which vapour and clusters are scavenged by pre-existing particles) to the growth rate of the clusters³. In the real world, both of

these quantities depend on the particle-size distribution.

The condensation sink can be derived directly from the particle-size distribution. However, the growth rate is commonly determined by monitoring how clusters grow over time, typically in the size range between 1 and 10 nanometres. This method assumes that the environmental factors that affect cluster growth are uniform throughout a given

region, and it has worked well in describing particle-growth behaviour in rural environments. However, it has failed to explain particle growth in cities⁵.

The particle loading of air in urban environments can be greater than 500 micrograms per cubic metre (ref. 6), whereas that of rural or remote environments is usually less than $5 \mu\text{g m}^{-3}$ (ref. 7). Newly formed clusters in cities must therefore rapidly scavenge vapour or combine with other clusters so that they can grow large enough for the rates at which they are themselves scavenged to be reduced (Fig. 1a), and therefore survive to become more-persistent, larger particles. Given that observed growth rates in urban areas are only a few times greater than those in remote environments, it is hard to understand how newly formed particles can reach diameters of 10 nm or more in urban areas – but such growth seems to be widespread in megacities in wintertime.

Wang *et al.* investigated this issue by carrying out a set of chamber experiments that reproduced atmospheric conditions typical of a megacity, focusing on the behaviour of ammonium nitrate. This compound is a crucial component of urban winter- and springtime particulate matter⁸, but has not been thought to have a major role in particle formation.

Ammonium nitrate exists in a temperature-dependent equilibrium with gaseous ammonia and nitric acid, and this equilibrium favours the gas phase when it is warm. However, the authors observed that ammonium nitrate rapidly condenses onto newly formed clusters at temperatures below 5 °C (Fig. 1b). This is

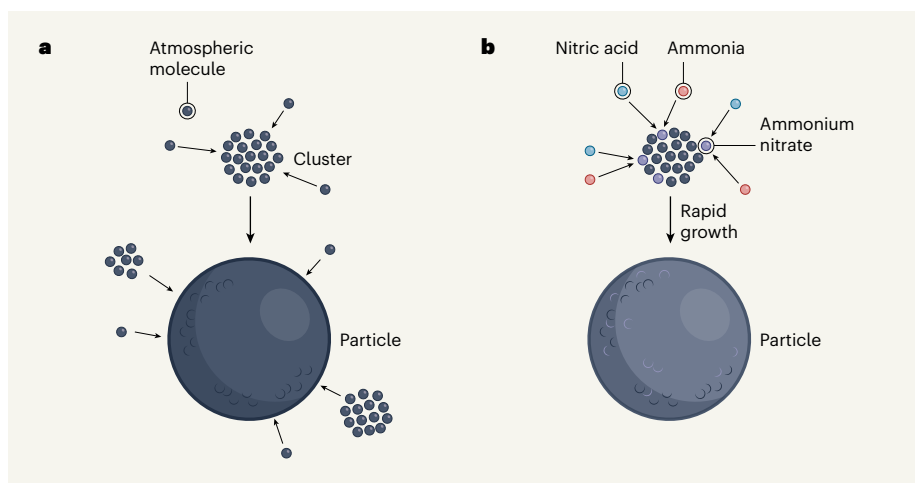


Figure 1 | The growth and formation of atmospheric particles. **a**, Small clusters of atmospheric molecules can gradually accumulate more molecules until they form stable particles. However, other particles in the atmosphere can scavenge the available vapour, limiting cluster growth, or even scavenge whole clusters. The concentration of particles in urban environments is high, which means that any clusters or vapour would be expected to be scavenged by existing particles before they form stable particles themselves. Yet the observed rate of new-particle formation is surprisingly high in megacities. **b**, Wang *et al.*¹ report that clusters can grow rapidly by accumulating ammonium nitrate (which forms from ammonia and nitric acid molecules) under conditions known to occur in megacities in winter. This allows the clusters to reach stable particle sizes before they are scavenged by other particles – and might explain the high particle-formation rates observed in urban areas.



Figure 2 | Heavy smog in Delhi during winter 2019. Severe wintertime pollution events in megacities can produce atmospheric conditions similar to those reported by Wang *et al.*¹ to cause rapid growth of atmospheric particles.

because the atmospheric concentrations of ammonia and nitric acid vapours at these temperatures can exceed their equilibrium values. To put it another way, when the ratio of the concentration of these gases to their concentration at equilibrium (the saturation ratio) under the same environmental conditions is greater than 1, rapid condensation takes place.

Crucially, the observed rapid condensation accelerates particle growth. The particle growth rates at -10°C were 200 times faster than those at $+5^{\circ}\text{C}$, for the same gas concentrations of ammonia and nitric acid. The growth rates at cold temperatures are much higher than those previously derived from field observations in urban areas.

By measuring the composition of vapours and particles using an array of advanced mass spectrometers, Wang and colleagues showed that ammonium nitrate does not participate in particle formation at temperatures above -15°C . Particle formation instead proceeds through a well-recognized pathway involving ammonia and sulfuric acid⁹; rapid growth through ammonium nitrate condensation begins to occur once a threshold cluster size has been reached. However, the authors report that new particles can form directly from ammonium nitrate at temperatures below -15°C . The authors speculate that this process could occur in the humid air outflows at the top of convective tropical clouds.

The authors show that the critical size at which ammonium nitrate starts to induce rapid growth depends on the saturation ratio of the ammonia–nitric acid system.

Furthermore, once a particle has reached that size, it continues to grow rapidly because the equilibrium concentration of ammonia and nitric acid above ammonium nitrate is much lower for larger particles. This growth occurs in much the same way that a liquid cloud forms on particles called condensation nuclei, which grow rapidly as soon as the saturation ratio of water exceeds 1.

So, how representative of the real world are these experimental observations, and what do

“This work provides key knowledge that will inform air-quality policy as the chemical composition of urban atmospheres changes in the future.”

they tell us about real urban environments? The mixing ratios of ammonia and nitric acid in the experiments are typical of those of many urban environments and are often greatly exceeded in some megacities. Moreover, in many places, such as Beijing or Delhi (Fig. 2), severe air-pollution events involving high concentrations of ammonia and nitric acid occur often, mostly in wintertime when the daytime temperatures are at, or below, 5°C (see ref. 10, for example).

However, air must become supersaturated with ammonia and nitric acid before clusters can grow through ammonium nitrate

condensation. Wang *et al.* convincingly argue that localized supersaturation of these gases is likely in many cities because the environment is heterogeneous. For example, the emission sources vary widely, and the flow of emissions around buildings, in street canyons and as a result of traffic movement combine to generate substantial gradients of concentration. The temperature in cities also often varies by several degrees over distances of a few metres to a few tens of metres, because of direct heating or shadowing from buildings, and because different surfaces absorb and reflect heat differently. These temperature variations can alter the saturation ratio of ammonia and nitric acid sufficiently for rapid condensation to occur.

Wang and colleagues calculate that the rapid condensation of ammonia and nitric acid occurs on timescales of several minutes in their experiments. The temperature heterogeneities observed in cities are sustained for similar timescales across various distances, potentially allowing clusters to grow to more-stable sizes at which further mass can be added to grow the particles. In other words, the new findings might explain why the initial stages of particle growth can be so fast in cities. Previously calculated cluster-growth rates in cities were averaged over space and time, and therefore did not capture this heterogeneity.

It will be extremely challenging to demonstrate that rapid ammonium nitrate condensation occurs in the real atmosphere, but the concept is very persuasive. Numerous semi-volatile organic compounds in the atmosphere might well have a similar role in particle growth. More broadly, Wang and colleagues' work provides key knowledge that will inform air-quality policy as the chemical composition of urban atmospheres changes in the future. Most notably, sulfur dioxide emissions are being reduced across many cities. This makes it increasingly likely that urban pollution will be dominated by emissions of nitrogen oxide (a precursor of nitric acid) from road traffic and by ammonia from agriculture for the coming decade or more.

Hugh Coe is in the Department of Earth and Environmental Sciences, University of Manchester, Manchester M13 9PL, UK.
e-mail: hugh.coe@manchester.ac.uk

1. Wang, M. *et al.* *Nature* **581**, 184–189 (2020).
2. Seinfeld, J. H. *Proc. Natl Acad. Sci. USA* **113**, 5781–5790 (2016).
3. Yao, L. *et al.* *Science* **361**, 278–281 (2018).
4. Ehn, M. *et al.* *Nature* **506**, 476–479 (2014).
5. Kulmala, M., Kerminen, V.-M., Petäjä, T., Ding, A. J. & Wang, L. *Faraday Discuss.* **200**, 271–288 (2017).
6. Chen, Y. *et al.* *Atmos. Environ.* **X** **5**, 100052 (2020).
7. Van Dingenen, R. *et al.* *Atmos. Environ.* **38**, 2561–2577 (2004).
8. Young, D. E. *et al.* *Atmos. Chem. Phys.* **15**, 6351–6366 (2015).
9. Kirkby, J. *et al.* *Nature* **476**, 429–433 (2011).
10. Huang, R.-J. *et al.* *Nature* **514**, 218–222 (2014).

Very regular high-frequency pulsation modes in young intermediate-mass stars

<https://doi.org/10.1038/s41586-020-2226-8>

Received: 17 July 2019

Accepted: 27 February 2020

Published online: 13 May 2020

 Check for updates

Timothy R. Bedding^{1,2✉}, Simon J. Murphy^{1,2}, Daniel R. Hey^{1,2}, Daniel Huber³, Tanda Li^{1,2,4}, Barry Smalley⁵, Dennis Stello^{2,6}, Timothy R. White^{1,2,7}, Warrick H. Ball^{2,4}, William J. Chaplin^{2,4}, Isabel L. Colman^{1,2}, Jim Fuller⁸, Eric Gaidos⁹, Daniel R. Harbeck¹⁰, J. J. Hermes¹¹, Daniel L. Holdsworth¹², Gang Li^{1,2}, Yaguang Li^{1,2,13}, Andrew W. Mann¹⁴, Daniel R. Reese¹⁵, Sanjay Sekaran¹⁶, Jie Yu¹⁷, Victoria Antoci^{2,18}, Christoph Bergmann⁶, Timothy M. Brown¹⁰, Andrew W. Howard⁸, Michael J. Ireland⁷, Howard Isaacson¹⁹, Jon M. Jenkins²⁰, Hans Kjeldsen^{2,21}, Curtis McCully¹⁰, Markus Rabus^{10,22}, Adam D. Rains⁷, George R. Ricker^{23,24}, Christopher G. Tinney⁶ & Roland K. Vanderspek^{23,24}

Asteroseismology probes the internal structures of stars by using their natural pulsation frequencies¹. It relies on identifying sequences of pulsation modes that can be compared with theoretical models, which has been done successfully for many classes of pulsators, including low-mass solar-type stars², red giants³, high-mass stars⁴ and white dwarfs⁵. However, a large group of pulsating stars of intermediate mass—the so-called δ Scuti stars—have rich pulsation spectra for which systematic mode identification has not hitherto been possible^{6,7}. This arises because only a seemingly random subset of possible modes are excited and because rapid rotation tends to spoil regular patterns^{8–10}. Here we report the detection of remarkably regular sequences of high-frequency pulsation modes in 60 intermediate-mass main-sequence stars, which enables definitive mode identification. The space motions of some of these stars indicate that they are members of known associations of young stars, as confirmed by modelling of their pulsation spectra.

The δ Scuti variables are stars of intermediate mass (1.5–2.5 solar masses, M_{\odot}) that pulsate in low-order pressure modes^{6,7}. Observations have shown that many δ Scuti stars have regular frequency spacings in their pulsation spectra (see Methods) but a large sample with unambiguous mode identifications is lacking. Each pulsation mode in a non-rotating star is identified by two integers: the radial order, n , and the degree, l . We expect the strongest observable modes to be of low degree ($l = 0, 1$ and 2), because higher degrees have greatly reduced amplitudes due to cancellation in disk-integrated light. In the so-called asymptotic regime ($n \gg l$), modes with a given degree l are approximately equally spaced in frequency by a separation, $\Delta\nu$, that is the inverse of the time taken for sound waves to travel through the star and is approximately proportional to the square root of the mean stellar density¹.

The patterns are more complex in a rotating star, with the mode frequencies also depending on the azimuthal order, m . Each non-radial ($l \geq 1$) mode in the pulsation spectrum is split into $2l + 1$ components,

where m ranges from $-l$ to l . The relative amplitudes of these components depend on the inclination of the rotation axis to the line of sight. For example, if a star is seen at low inclination (close to pole-on) then the axisymmetric ($m = 0$) mode in each multiplet will dominate, leading to a simpler pulsation spectrum. In very rapidly rotating stars, the oblateness alters the pulsation cavity and further complicates the pattern. However, for rotation rates less than about 50% of Keplerian break-up, the radial modes ($l = 0$) and the axisymmetric dipolar modes ($l = 1, m = 0$) are still expected¹¹ to follow a regular spacing that is similar to the non-rotating case, but with a slightly smaller $\Delta\nu$.

To search for regular patterns we have used observations from the Transiting Exoplanet Survey Satellite (TESS), which provides light curves for many thousands of δ Scuti stars at rapid cadence (120-s sampling). We used the first nine 27-day sectors of TESS data and focused on identifying δ Scuti stars that pulsate at high frequencies (above about 30 d^{−1}). We also examined stars not previously known to pulsate by calculating the Fourier spectra of TESS light curves and measuring

¹Sydney Institute for Astronomy (SIfA), School of Physics, University of Sydney, Camperdown, New South Wales, Australia. ²Stellar Astrophysics Centre, Department of Physics and Astronomy, Aarhus University, Aarhus, Denmark. ³Institute for Astronomy, University of Hawai'i, Honolulu, HI, USA. ⁴School of Physics and Astronomy, University of Birmingham, Birmingham, UK.

⁵Astrophysics Group, Lennard-Jones Laboratories, Keele University, Keele, UK. ⁶School of Physics, University of New South Wales, Kensington, New South Wales, Australia. ⁷Research School of Astronomy and Astrophysics, Mount Stromlo Observatory, The Australian National University, Canberra, Australian Capital Territory, Australia. ⁸TAPIR, California Institute of Technology, Pasadena, CA, USA. ⁹Department of Earth Sciences, University of Hawai'i, Honolulu, HI, USA. ¹⁰Las Cumbres Observatory Global Telescope, Goleta, CA, USA. ¹¹Department of Astronomy, Boston University, Boston, MA, USA. ¹²Jeremiah Horrocks Institute, University of Central Lancashire, Preston, UK. ¹³Department of Astronomy, Beijing Normal University, Beijing, China. ¹⁴Department of Physics and Astronomy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. ¹⁵LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université de Paris, Meudon, France. ¹⁶Instituut voor Sterrenkunde (IVS), KU Leuven, Leuven, Belgium. ¹⁷Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany. ¹⁸DTU Space, National Space Institute, Technical University of Denmark, Kongens Lyngby, Denmark. ¹⁹Department of Astronomy, University of California at Berkeley, Berkeley, CA, USA. ²⁰NASA Ames Research Center, Moffett Field, CA, USA. ²¹Institute of Theoretical Physics and Astronomy, Vilnius University, Vilnius, Lithuania. ²²Department of Physics, University of California, Santa Barbara, CA, USA. ²³Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. ²⁴Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA, USA.

✉e-mail: tim.bedding@sydney.edu.au

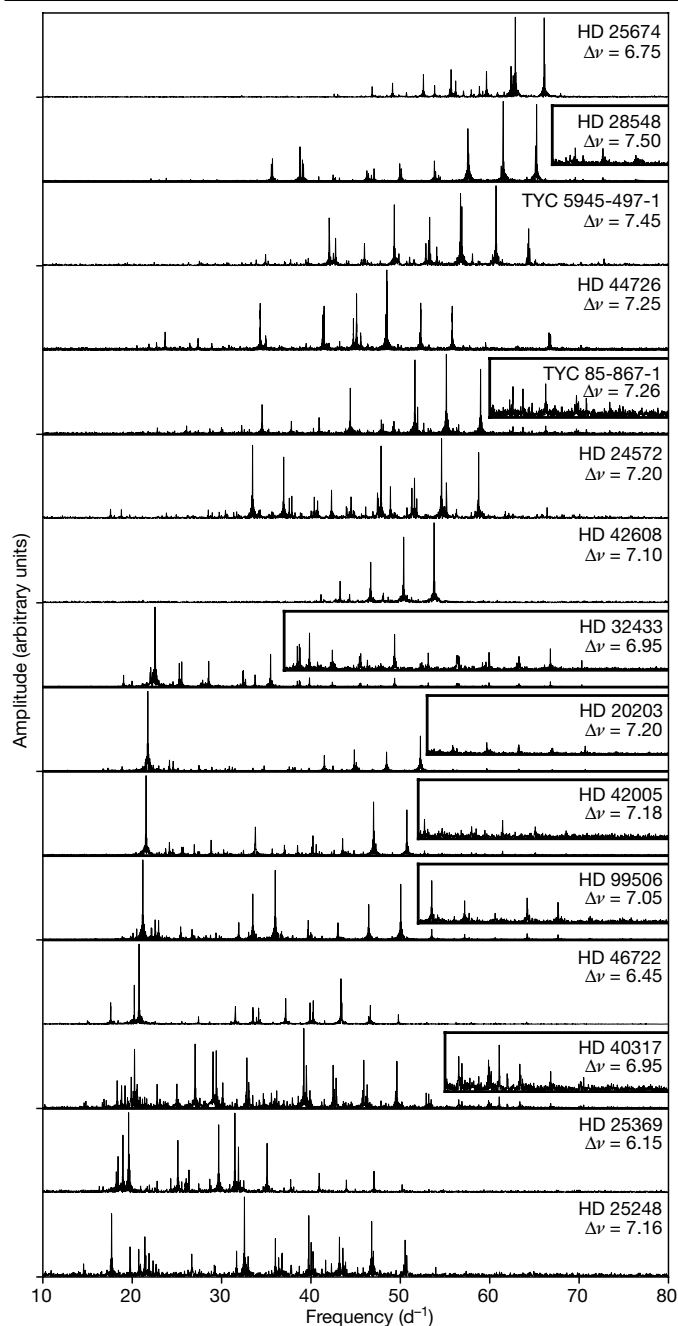


Fig. 1 | Pulsation spectra of 15 high-frequency δ Scuti stars observed with TESS. The measured value of $\Delta\nu$ (in d^{-1}) is given in each panel (see Extended Data Table 1). Insets for some spectra expand the vertical axis by a factor of four to make weaker peaks more visible.

the skewness of the distribution of peak heights¹² above 30 d^{-1} as a way to flag likely detections.

We then inspected the pulsation spectra for regularity using échelle diagrams (described below). In addition, we used data from the Kepler spacecraft, which observed about 300 δ Scuti stars at short cadence (60-s sampling) during its four-year nominal mission^{12–14}. Stars observed in Kepler’s long-cadence mode (29.4-min sampling) were not considered because the Nyquist frequency of 24.5 d^{-1} makes it very difficult to identify patterns in high-frequency pulsators.

We discovered 60 stars with regular frequency spacings (Extended Data Table 1), which define a group of δ Scuti stars for which mode identification is possible. Fig. 1 shows some of the pulsation spectra,

which have remarkably regular patterns of peaks. The small amplitudes of the highest-frequency modes may indicate that turbulent pressure, rather than the standard opacity mechanism, is responsible for driving them¹⁵. About one-third of the stars in our sample (for example, the bottom half of Fig. 1) show a strong peak in the range $18\text{--}23 \text{ d}^{-1}$, which is likely to be the fundamental radial pressure mode ($n=1, l=0$). This identification is strengthened by the fact that these peaks agree with the established period–luminosity relation for the fundamental radial mode in δ Scuti stars¹⁶, and by the fact that we find a good correlation between this frequency (when present) and the measured value of $\Delta\nu$ (Extended Data Fig. 2). In addition, six stars show a mode that is a factor of about 0.78 shorter in period, consistent with being the first radial overtone ($n=2, l=0$)¹⁷.

Fig. 2 shows the pulsation spectra of several δ Scuti stars in échelle format, where the spectra have been divided into equal segments of width $\Delta\nu$ and stacked vertically so that peaks with the same degree fall along vertical ridges. The regularity of the patterns is striking, similar to échelle diagrams of solar-like oscillators^{1–3} but at much lower radial orders. Comparison with pulsation frequencies calculated from theoretical models (red symbols in Fig. 2a–c) enables an unambiguous identification of ridges corresponding to sequences of radial modes ($l=0$) and dipolar modes ($l=1$), as shown (more examples are shown in Extended Data Fig. 1). Sequences of $l=2$ modes do not appear to be present in these stars.

We have placed our sample in the Hertzsprung–Russell (H–R) diagram using effective temperatures and luminosities derived from broadband colours and Gaia parallaxes (Fig. 3a). The δ Scuti stars with regular frequency spacings tend to be located near the zero-age main sequence (ZAMS), with masses between $1.5M_{\odot}$ and $1.8M_{\odot}$. The fact that these stars are relatively young helps to explain their regular pulsation spectra. In more evolved stars, the non-radial modes are expected to be ‘bumped’ from their regular spacings when they undergo avoided crossings due to coupling with gravity (buoyancy) modes in the core³. For young stars, this mode bumping only occurs at the lowest frequencies, as can be seen from the models of $l=1$ modes in Fig. 2a–c (red triangles at low frequencies).

The large frequency separation, $\Delta\nu$, scales approximately as the square root of the mean stellar density^{11,18–20}. However, the mode spacings of stars are not completely regular—even in the asymptotic regime—meaning that $\Delta\nu$ varies with frequency. We used theoretical models to calculate $\Delta\nu$ for δ Scuti stars in the same region that we measured it, namely from radial modes with orders in the range $n=4$ to 8 (see Methods). We found that $\Delta\nu$ in the models was typically 15% lower than would be obtained by scaling from the density of the Sun, which is consistent with previous results^{10,18–20}. Fig. 3b compares the observed large separations of our sample with the densities derived from fitting to evolutionary tracks in the H–R diagram. The results confirm there is a correlation, with most stars lying between the values based on the standard scaling relation (solid red curve) and those from the model calculations (dashed red curve). Some of the spread is probably due to the range of metallicities of the sample, and some will be due to rotation. For example, if a star is oblate due to rotation then the mean density will be reduced. In addition, the inclination of its rotation axis affects the observed position in the H–R diagram²¹ (and hence the inferred radius, mass and density). The absolute position of the regular comb pattern, parametrized by the phase term ε (see Methods), also contains important information about the interior structure of the star. In solar-type stars, the value of ε does not change greatly during evolution²². In these intermediate-mass stars, this appears not to be the case and ε serves as a useful indicator for age (Fig. 3c).

High-resolution spectroscopy can be used to measure $v \sin i$, the projected rotational velocity of a star (where v is the equatorial velocity and i is the inclination angle), and most intermediate-mass stars have $v \sin i$ values²³ in the range $50\text{--}220 \text{ km s}^{-1}$. Measurements are available for 39 of the 60 stars in our sample (see Extended Data Table 2), of

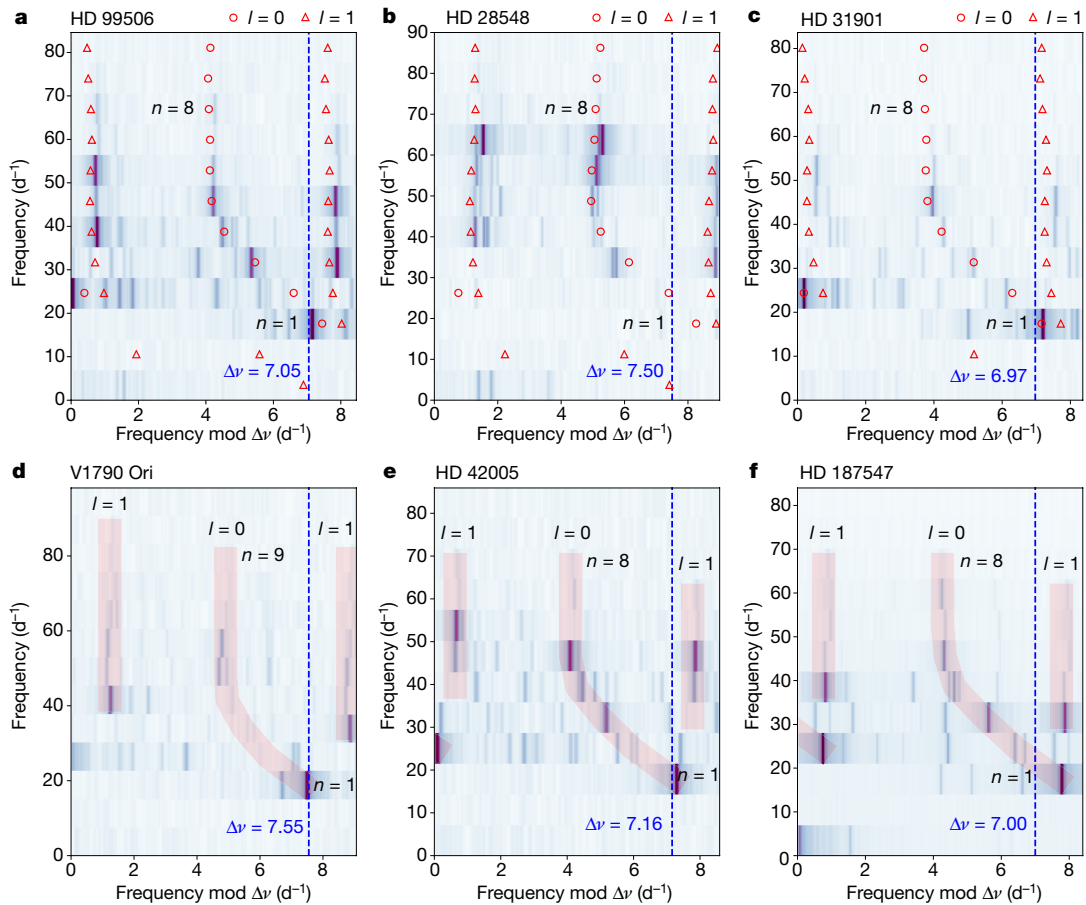


Fig. 2 | Mode identification in δ Scuti stars. a–f, Pulsation spectra of HD 99506 (a), HD 28548 (b), HD 31901 (c), V1790 Ori (d), HD 42005 (e) and HD 187547 (f). Spectra are shown in échelle format, segments of equal length being stacked vertically. In each panel, the vertical dashed line shows the value of $\Delta\nu$, with a repeated overlap region added on the right for clarity. The greyscale shows the observed amplitude spectrum, which in most cases was calculated from one 27-day sector of data from the TESS spacecraft. The exception is HD 187547, for which observations were made over 960 d with the Kepler spacecraft³². Some smoothing was applied to the observed amplitude spectra before plotting. In the top row (a–c), the red symbols show mode frequencies calculated from theoretical models of non-rotating stars, chosen

to match the observed modes reasonably well (see Methods). These allow mode identifications in other stars, as shown in the bottom row (d–f), where the red stripes mark overtone sequences of $l=0$ and $l=1$ modes. The parameters of the models shown in a–c are as follows (while noting that other values of the parameters also give fits of similar quality): **a**, HD 99506: mass $1.68M_{\odot}$, metallicity $[\text{Fe}/\text{H}] = 0.0$, age 200 Myr, effective temperature 8,065 K and radius $1.51R_{\odot}$. **b**, HD 28548: mass $1.59M_{\odot}$, metallicity $[\text{Fe}/\text{H}] = -0.2$, age 270 Myr, effective temperature 8,202 K and radius $1.41R_{\odot}$. **c**, HD 31901: mass $1.77M_{\odot}$, metallicity $[\text{Fe}/\text{H}] = 0.08$, age 102 Myr, effective temperature 8,083 K and radius $1.51R_{\odot}$.

which 17 stars have $v \sin i \leq 50 \text{ km s}^{-1}$. Thus, our sample of δ Scuti stars includes many with unusually low projected rotational velocities, which is consistent with the idea that regular frequency spacings are more common in stars seen at high inclinations (close to pole-on).

Some échelle diagrams show the modes along the $l=1$ ridge to be split into close doublets, as expected for rotating stars (some examples, namely HD 24975 and HD 46722, are shown in Extended Data Fig. 1). Four échelle diagrams show more complicated patterns, with additional ridges at various angles that indicate sequences with slightly different spacings (Fig. 4). The rotation axes of these stars are presumably at higher inclinations than those with simpler pulsation spectra, which would lead us to expect one $l=0$ ridge and three $l=1$ ridges. Beyond the usual rotational splitting of $l=1$ modes, slightly different frequency spacings are expected for each m in an oblate star. This is because modes with different m propagate along different paths through the star, giving different values for the sound-speed crossing time and hence for $\Delta\nu$. In stars with even more ridges, the additional sequences could correspond to modes with higher degrees ($l \geq 2$)—where coupling between modes with different degree may also be important—and perhaps also to chaotic modes^{9,24}.

The identification of regular pulsation frequency patterns in intermediate-mass stars will expand the reach of asteroseismology to new frontiers. One example is to determine the ages of young moving groups, clusters and stellar streams, which can vary by a factor of up to two, depending on the method used²⁵. Spectroscopic radial velocities and Gaia astrometry show that several stars in our sample are members of nearby young associations (references given in Extended Data Table 1), including the Octans association (HD 44930, HD 29783, HD 42915), the Carina association (HD 89263), the Columba association (HD 37286 = HR 1915), the β Pictoris moving group (β Pic itself) and the recently discovered Pisces–Eridanus stellar stream (HD 31901). For the last, gyrochronology yielded an age similar to that of the Pleiades (about 130 Myr)²⁶, in contrast to the initial approximately 1-Gyr age determination from suspected evolved moving group members²⁷. Asteroseismic modelling of HD 31901 (Fig. 2c) clearly confirms a young age for this member of the Pisces–Eridanus group (see Methods), and similar age determinations might be possible for other groups containing intermediate-mass stars.

Four stars in our sample (HD 28548, HD 34282, TYC 5945-497-1 and V1790 Ori) exhibit excess emission in the WISE passbands, indicating

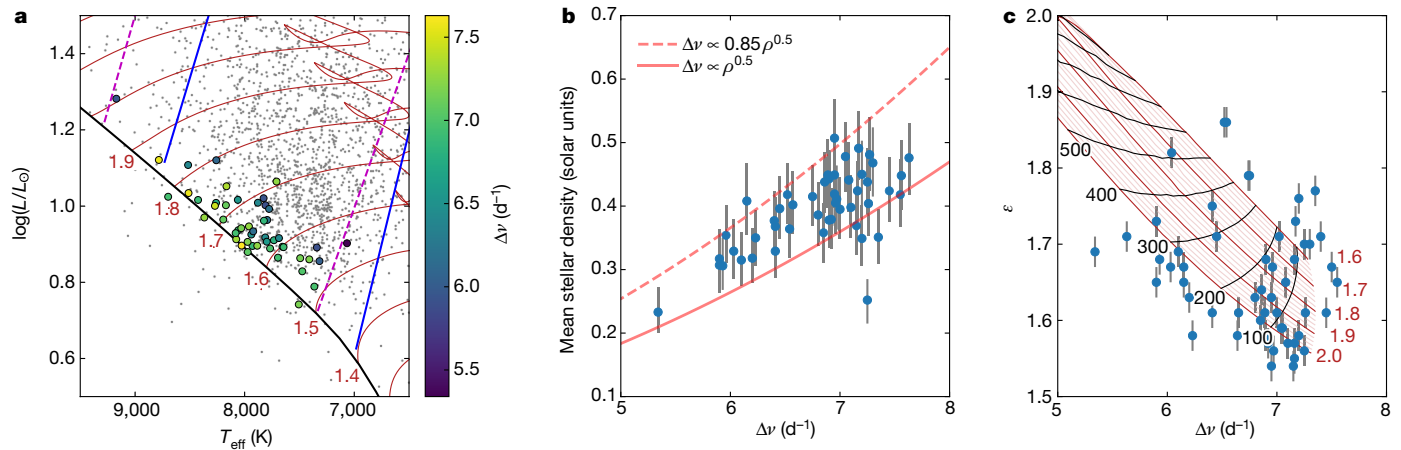


Fig. 3 | Properties of high-frequency δ Scuti stars. **a**, Location of our sample in the H–R diagram (filled circles, colour-coded by the measured large frequency separation $\Delta\nu$; L , luminosity; T_{eff} , effective temperature). The small points show δ Scuti stars observed by the Kepler mission¹² and the red curves (labelled by mass in solar units) are evolutionary tracks calculated for solar metallicity (see Methods). The solid blue lines show the edges of the theoretical δ Scuti instability strip, and the dashed magenta lines show the observed instability strip based on Kepler stars¹². **b**, Mean stellar density, ρ , versus large frequency separation as determined from observations (symbols; error bars, 1σ uncertainties), as predicted from the standard scaling relation (solid red line) and from non-rotating stellar models (red dashed line). Stars with close

binary companions have been omitted from **a** and **b** (see Methods). **c**, The phase term ε , which measures the absolute position of the oscillation spectrum, versus large frequency separation (symbols; error bars, 1σ uncertainties). Red curves (labelled by mass in solar units) are evolutionary tracks based on fitting to radial modes with $n = 4$ to 8 (see Methods), and shorter black curves are the corresponding isochrones, labelled in Myr. These models are only intended to be indicative, since they are calculated for solar metallicity and do not include rotation, which affects both $\Delta\nu$ and ε . The models do show that, unlike for solar-type stars²², ε varies substantially during the evolution and is therefore sensitive to age, which is an important bonus for asteroseismology of δ Scuti stars.

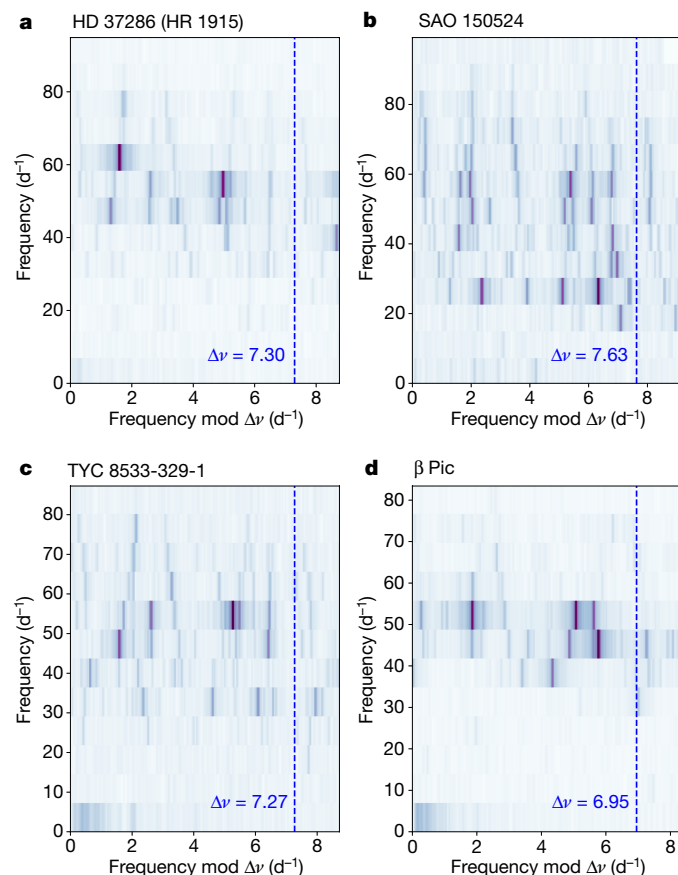


Fig. 4 | Examples of more complicated échelle diagrams of δ Scuti pulsations. **a–d**, Diagrams for HD 37286 (=HR 1915; **a**), SAO 150524 (**b**), TYC 8533-329-1 (**c**) and β Pic (**d**). There are sets of ridges at a range of angles, indicating slightly different spacings. An intermediate value of $\Delta\nu$ was chosen for these diagrams (see Methods).

a circumstellar dust disk. One of these (HD 34282) has a disk that has been resolved by ALMA, showing it to be inclined $60^\circ \pm 1^\circ$ to the line of sight²⁸. The constraints on age and inclination of this host star provided by an analysis of its pulsations could illuminate the origin of stellar obliquity²⁹ and the pace of disk evolution³⁰.

Finally, we note that six stars in our sample have been classified spectroscopically as λ Boötis stars (references given in Extended Data Table 1), meaning that their surface chemical abundances show evidence for accretion from circumstellar material. Given that λ Boötis stars are rare, making up only about 2% of A stars³¹, the relatively high occurrence rate in our sample lends support to the hypothesis that λ Boötis stars tend to be young, with circumstellar material accreting from a proto-planetary disk.

The stars observed by TESS at 2-min cadence constitute a small fraction of stars that fall on the full-frame images (FFIs). Future TESS observations should reveal many more examples of δ Scuti stars with high-frequency overtones, especially given that the cadence of TESS FFIs will switch from 30 min to 10 min in the extended mission that starts in July 2020. It is likely that the stars with regular patterns can guide mode identification in the much larger number of δ Scuti stars whose pulsation spectra are not as regular.

Online content

Any methods, additional references, Nature Research reporting summaries, articles, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2226-8>.

1. Aerts, C., Christensen-Dalsgaard, J. & Kurtz, D. W. *Asteroseismology* (Springer, 2010).
2. García, R. A. & Ballot, J. Asteroseismology of solar-type stars. *Living Rev. Phys. Sol.* **16**, 4 (2019).
3. Hekker, S. & Christensen-Dalsgaard, J. Giant star seismology. *Astron. Astrophys. Rev.* **25**, 1 (2017).
4. Aerts, C. in *New Windows on Massive Stars: Asteroseismology, Interferometry, and Spectropolarimetry* (eds Meynet, G., Georgy, C., Groh, J. & Stee, P.) 154–164 (IAU Symp. 307, Cambridge Univ. Press, 2015).

5. Córscico, A. H., Althaus, L. G., Miller Bertolami, M. M. & Kepler, S. O. Pulsating white dwarfs: new insights. *Astron. Astrophys. Rev.* **27**, 7 (2019).
6. Goupil, M. J. et al. Asteroseismology of δ Scuti stars: problems and prospects. *J. Astron. Astrophys.* **26**, 249–259 (2005).
7. Handler, G. Delta Scuti variables. *AIP Conf. Proc.* **1170**, 403–409 (2009).
8. Ouazzani, R. M., Roxburgh, I. W. & Dupret, M. A. Pulsations of rapidly rotating stars. II. Realistic modelling for intermediate-mass stars. *Astron. Astrophys.* **579**, A116 (2015).
9. Reese, D. R. et al. Frequency regularities of acoustic modes and multi-colour mode identification in rapidly rotating stars. *Astron. Astrophys.* **601**, A130 (2017).
10. Mirouh, G. M., Angelou, G. C., Reese, D. R. & Costa, G. Mode classification in fast-rotating stars using a convolutional neural network: model-based regular patterns in δ Scuti stars. *Mon. Not. R. Astron. Soc.* **483**, L28–L32 (2019).
11. Reese, D., Lignières, F. & Rieutord, M. Regular patterns in the acoustic spectrum of rapidly rotating stars. *Astron. Astrophys.* **481**, 449–452 (2008).
12. Murphy, S. J., Hey, D., Van Reeth, T. & Bedding, T. R. Gaia-derived luminosities of Kepler A/F stars and the pulsator fraction across the δ Scuti instability strip. *Mon. Not. R. Astron. Soc.* **485**, 2380–2400 (2019).
13. Balona, L. A., Daszyńska-Daszkiewicz, J. & Pamyatnykh, A. A. Pulsation frequency distribution in δ Scuti stars. *Mon. Not. R. Astron. Soc.* **452**, 3073–3084 (2015).
14. Bowman, D. M. & Kurtz, D. W. Characterizing the observational properties of δ Sct stars in the era of space photometry from the Kepler mission. *Mon. Not. R. Astron. Soc.* **476**, 3169–3184 (2018).
15. Antoci, V. et al. The first view of δ Scuti and γ Doradus stars with the TESS mission. *Mon. Not. R. Astron. Soc.* **490**, 4040–4059 (2019).
16. Ziaali, E., Bedding, T. R., Murphy, S. J., Van Reeth, T. & Hey, D. R. The period-luminosity relation for δ Scuti stars using Gaia DR2 parallaxes. *Mon. Not. R. Astron. Soc.* **486**, 4348–4353 (2019).
17. Petersen, J. O. & Christensen-Dalsgaard, J. Pulsation models of δ Scuti variables. I. The high-amplitude double-mode stars. *Astron. Astrophys.* **312**, 463–474 (1996).
18. Suárez, J. C. et al. Measuring mean densities of δ Scuti stars with asteroseismology. Theoretical properties of large separations using TOUCAN. *Astron. Astrophys.* **563**, A7 (2014).
19. García Hernández, A. et al. Observational $\Delta\nu$ - ρ relation for δ Sct stars using eclipsing binaries and space photometry. *Astrophys. J.* **811**, L29 (2015).
20. Paparó, M., Benkő, J. M., Hareter, M. & Guzik, J. A. Unexpected series of regular frequency spacing of δ Scuti stars in the non-asymptotic regime. II. Sample-échelle diagrams and rotation. *Astrophys. J. Suppl. Ser.* **224**, 41 (2016).
21. Suárez, J. C. et al. A study of correlation between the oscillation amplitude and stellar parameters of δ Scuti stars in open clusters. Toward selection rules for δ Scuti star oscillations. *Astron. Astrophys.* **390**, 523–531 (2002).
22. White, T. R. et al. Calculating asteroseismic diagrams for solar-like oscillations. *Astrophys. J.* **743**, 161 (2011).
23. Zorec, J. & Royer, F. Rotational velocities of A-type stars. IV. Evolution of rotational velocities. *Astron. Astrophys.* **537**, A120 (2012).
24. Evano, B., Lignières, F. & Georgeot, B. Regularities in the spectrum of chaotic p-modes in rapidly rotating stars. *Astron. Astrophys.* **631**, A140 (2019).
25. Mamajek, E. E. & Bell, C. P. M. On the age of the β Pictoris moving group. *Mon. Not. R. Astron. Soc.* **445**, 2169–2180 (2014).
26. Curtis, J. L., Agüeros, M. A., Mamajek, E. E., Wright, J. T. & Cummings, J. D. TESS reveals that the nearby Pisces-Eridanus stellar stream is only 120 Myr old. *Astron. J.* **158**, 77 (2019).
27. Meingast, S., Alves, J. & Fürnkranz, V. Extended stellar systems in the solar neighborhood. II. Discovery of a nearby 120° stellar stream in Gaia DR2. *Astron. Astrophys.* **622**, L13 (2019).
28. van der Plas, G. et al. An 80 au cavity in the disk around HD 34282. *Astron. Astrophys.* **607**, A55 (2017).
29. Lai, D. Star-disc-binary interactions in protoplanetary disc systems and primordial spin-orbit misalignments. *Mon. Not. R. Astron. Soc.* **440**, 3532–3544 (2014).
30. Williams, J. P. & Cieza, L. A. Protoplanetary disks and their evolution. *Annu. Rev. Astron. Astrophys.* **49**, 67–117 (2011).
31. Paunzen, E. A spectroscopic survey for λ Bootis stars. III. Final results. *Astron. Astrophys.* **373**, 633–640 (2001).
32. Antoci, V. et al. The role of turbulent pressure as a coherent pulsational driving mechanism: the case of the δ Scuti star HD 187547. *Astrophys. J.* **796**, 118 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Pulsation analysis

Light curves from TESS³³ and Kepler³⁴ were downloaded from MAST (Barbara A. Mikulski Archive for Space Telescopes)³⁵. We used Pre-search Data Conditioning Simple Aperture Photometry (PDCSAP) to calculate the Fourier amplitude spectra using a standard Lomb–Scargle periodogram.

For TESS, we examined all 92,000 stars having 2-min light curves in sectors 1–9. We used the skewness of the distribution of peak heights¹² above 30 d^{−1} as a way to identify high-frequency δ Scuti pulsators, producing a list of about 1,000 stars. Inspecting their échelle diagrams (see below) revealed 57 δ Scuti stars having a regular series of high-frequency peaks. For Kepler, we looked at all (about 330) δ Scuti stars that have short-cadence data (60-s sampling) and identified three stars with regular peaks.

The large separations ($\Delta\nu$) for the 60 stars in our sample are listed in Extended Data Table 1. In most cases, $\Delta\nu$ was measured by aligning the highest-frequency radial modes in a vertical ridge in the échelle diagram using the Python package *echelle*³⁶, which allows the value of $\Delta\nu$ to be fine-tuned interactively. This allowed $\Delta\nu$ to be measured to a precision of about 0.02 d^{−1} (see examples in Fig. 2 and Extended Data Fig. 1). Four stars do not show a clear sequence of radial modes, with the échelle diagrams showing several ridges that are not quite parallel (Fig. 4). In these cases, we chose $\Delta\nu$ to be the average of the values needed to make the individual ridges vertical.

The phase term ε is given for those stars having a clear $l = 0$ sequence, as determined from the horizontal position of that ridge in the échelle diagram. Note that $\Delta\nu$ and ε are related to the frequencies of high-order radial modes via the asymptotic relation^{1–3}: $\nu_{n,l=0} \approx \Delta\nu(n + \varepsilon)$. The uncertainty in ε determined in this way is about 0.02.

To rule out contamination from nearby stars as the source of the observed pulsations, we examined the pixel data and cross-matched with the Gaia DR2 catalogue. We considered a region of 5 × 5 TESS pixels (63 × 63 arcsec) centred on each target. We found that no dilution is present in one-third of the targets, with most of the remainder having small amounts of dilution (0.1%–3%). Only five stars have dilutions above 8%. We conclude that contamination of the photometry from nearby stars is negligible.

Fundamental stellar properties

To estimate properties for our sample we used Tycho B_T and V_T photometry³⁷, which we transformed into Johnson B and V magnitudes³⁸. We then used a $(B - V) - T_{\text{eff}}$ relation³⁹, Gaia DR2 parallaxes⁴⁰, a 3D dust map⁴¹, and V-band bolometric corrections to calculate effective temperatures and luminosities. We did this by solving for the distance modulus, as implemented in the ‘direct mode’ version of isoclassify⁴². For stars with typical uncertainties >0.01 mag in Tycho ($V_T > 9$ mag), we used the Gaia BP – RP colour index (with which we interpolated the colour – T_{eff} relation in the MIST (MESA Isochrones and Stellar Tracks) model grid⁴³ for solar metallicity) to derive T_{eff} , and we used 2MASS K-band magnitudes in combination with Gaia parallaxes to derive luminosities.

We adopted 2% fractional uncertainties for all derived effective temperatures, which is typical of the residual scatter in optical colour–temperature relations⁴⁴. A comparison of our Gaia-derived temperatures with those derived from Tycho photometry for stars with $V_T < 10$ mag, and a comparison with an independent implementation of the infrared flux method (IRFM), both showed good agreement with no systematic offsets. Our effective temperatures are on average about 1.5% (200 K) hotter than those for A-type stars in the Kepler Stellar Properties Catalog^{45,46}, which were predominantly based on the Kepler Input Catalog (KIC)⁴⁷. Such systematic differences are typical for effective temperature scales in A stars, reflecting the fact that the KIC was not optimized for A stars.

To estimate mean stellar densities, we fitted the effective temperatures and luminosities derived in the previous step to MIST isochrones

using the ‘grid mode’ of isoclassify, assuming a solar-neighbourhood metallicity prior. The procedure also yielded estimates of stellar masses and surface gravities, which combined with T_{eff} were used for the interpolation of bolometric corrections in the previous step. We iterated between the ‘direct mode’ and ‘grid mode’ calculations until all values converged, and adopted 0.03 mag bandpass-independent uncertainties in reddening and bolometric corrections. Extended Data Table 1 lists all stellar properties of the sample. Typical uncertainties are about 5% in luminosity and about 15% in mean stellar density. The properties of V1366 Ori (HD 34282) are not shown because they are highly uncertain due to obscuration by circumstellar material (it is classified as a Herbig Ae star)⁴⁸. This star is not plotted in Fig. 3.

To identify close binaries, which could bias the derived stellar parameters, we cross-matched our targets with the Washington Double Star catalogue (WDS). We also calculated the Gaia DR2 re-normalized unit weight error (RUWE) for each target, which provides a quality metric that accounts for the effects of colour and apparent magnitude on Gaia astrometric solutions. Stars with WDS companions within 2 arcsec or Gaia RUWE > 2 do not have parameters in Extended Data Table 1 and were not plotted in Fig. 3.

High-resolution spectroscopy

We obtained optical high-dispersion spectra of some stars in the sample in April and May 2019 using the HIRES spectrograph⁴⁹ at the Keck-I 10-m telescope on Maunakea observatory, Hawai‘i. The spectra were obtained and reduced as part of the California Planet Search queue⁵⁰. We typically obtained 1-min integrations using the C5 decker, resulting in a signal to noise (S/N) per pixel of 50 at 600 nm with a spectral resolution of $R \approx 60,000$.

High-resolution spectra for some stars were obtained in May and June 2019 using the NRES spectrograph⁵¹ at the Las Cumbres Observatory Global Telescope Network⁵² 1-m telescopes at Cerro Tololo Inter-American Observatory, Chile, and at Sutherland, South Africa. Exposure times were typically 10 min, resulting in a S/N per resolution element above 70 at about 510 nm, with a spectral resolution of $R \approx 50,000$. High-resolution spectra for an additional nine stars were obtained in June 2019 using the Veloce Rosso spectrograph⁵³ at the 3.9-m Anglo-Australian Telescope (AAT). These spectra covered the range 580–930 nm at a resolution of $R \approx 75,000$. Typical exposure times were 5–10 min (in cloudy conditions), resulting in a S/N per pixel of 50–90 at about 780 nm.

Extended Data Fig. 3 shows a small region of some of these spectra, alongside the Fourier amplitude spectra. The spectral analysis was performed using the UCLSYN spectral synthesis package^{54,55} using ATLAS9 models without convective overshooting⁵⁶. Atomic data used in the analysis was obtained from the VALD database⁵⁷, using their default search and extraction parameters. Surface gravities were fixed to $\log g = 4.0$ for all stars in the analysis. A microturbulent velocity of $\xi = 3 \text{ km s}^{-1}$ was assumed, which is the typical value for stars within the spectral range considered here^{58,59}. Measurements of the projected equatorial rotation velocity ($v \sin i$) were obtained through individual fits to several small (5 nm) regions between 500 nm and 550 nm (and 620–650 nm plus 778 nm for the AAT spectra), avoiding any inter-order gaps. The final values were determined by calculating the mean and standard deviation of the values obtained in the small spectral regions.

Independent $v \sin i$ values were determined for five of the spectra using the Grid Search in Stellar Parameters (GSSP) software⁶⁰. GSSP is designed to fit a grid of synthetic spectra with varying T_{eff} , $\log g$, ξ , $v \sin i$ and $[M/H]$ to each observed spectrum and output the χ^2 values of the fit. These synthetic spectra are generated on-the-fly during the fitting process using the SYNTHV radiative transfer code⁶¹ combined with a grid of atmospheric models from the LLMODELS code⁶². We fixed the microturbulent velocity at $\xi = 2.0 \text{ km s}^{-1}$ to prevent degeneracies with metallicity. The derived values were found to agree within uncertainties with the results from the UCLSYN spectral synthesis.

For a further nine stars, we estimated $\nu \sin i$ using low-resolution spectra that were obtained either with the RSS instrument on the Southern African Large Telescope (SALT)^{63–65} or the ISIS instrument on the William Herschel Telescope (WHT). Exposure times were typically a few minutes, which provided a S/N of about 100 at a spectral resolution of $R \approx 3,000$. For each target, a coarse grid of synthetic models was constructed using the stellar parameters in Extended Data Table 1 and a range of $\nu \sin i$ values. The observations were compared to the synthetic spectra to estimate the $\nu \sin i$ and the associated uncertainty.

Extended Data Table 2 lists the determined $\nu \sin i$ values for each star. Values in parentheses indicate close binaries (see above), meaning that $\nu \sin i$ may not be reliable.

To determine membership of moving groups, clusters and stellar streams, we calculated barycentric radial velocities using the Python implementation `barycorrPy`⁶⁶ of the barycentric correction algorithm of Wright et al.⁶⁷. These were combined with space motions calculated from Gaia DR2 astrometry, and Bayesian posterior probabilities of membership in known nearby moving groups were calculated using `Banyan Σ` ⁶⁸.

Stellar models

The stellar models presented in Fig. 2 used the ‘astero’ extension of MESA (Modules for Experiments in Stellar Astrophysics)^{69–71}. We used two approaches that gave similar results. One was based on a model grid calculated with MESA (v8118), where we varied mass from $1.3M_{\odot}$ to $1.9M_{\odot}$ in steps of $0.01M_{\odot}$ and metallicity ($[Fe/H]$) from -0.5 to 0.5 in steps of 0.1 . We used a fixed (solar-calibrated) mixing-length parameter of $\alpha_{MLT} = 1.9$ and a helium-to-heavy-element enrichment ratio of 1.33 . The best-fitting model was found by maximum likelihood estimation, where we included effective temperature, metallicity, luminosity and all identified pulsation frequencies. Equal weight was given in the likelihood function to the following five observables: frequencies of radial modes, frequencies of dipolar modes, effective temperature, metallicity and luminosity. The other approach used the automated simplex search in MESA-astero (v7503), where the fit was guided by the observed radial modes only. The search was allowed to vary the mass, metallicity, mixing length, and the age of the model in order to converge to the best fit. A helium-to-heavy-element enrichment ratio of 1.4 was used. Both approaches assumed a primordial helium abundance of 0.249 and we did not make any correction for surface effects in the way that is commonly done for solar-like stars⁷².

For the three examples shown in the upper row of Fig. 2, the agreement between models and observations is sufficiently good that we can unambiguously identify the two sequences corresponding to $l = 0$ and $l = 1$ modes. One noteworthy feature of the models and the observations is that the $l = 0$ sequence bends to the right at the bottom of each figure, indicating that $\Delta \nu$ decreases towards the lowest-order modes, whereas the $l = 1$ sequence does not show this effect. This difference is a general feature of these models and makes it possible to identify the sequences in other stars, as shown in the lower row of Fig. 2 and in Extended Data Fig. 1.

For Fig. 3a we used the evolutionary tracks with solar metallicity ($X = 0.71, Z = 0.014$) from Murphy et al.¹². The other parameters of those tracks are $\alpha_{MLT} = 1.8$, exponential core overshooting of $0.015H_p$ (pressure scale heights), exponential over- and undershooting of $0.015H_p$ for the hydrogen-burning shell, exponential envelope overshooting of $0.025H_p$, diffusive mixing $\log(D_{mix}) = 0$ (with D_{mix} in $\text{cm}^2 \text{s}^{-1}$), OPAL opacities and the solar abundance mixture⁷³. As noted by Murphy et al.¹², these tracks are in good agreement with the MIST tracks computed with no rotation and similar metallicities, except that the latter have a shorter main-sequence phase. This is not expected to be important for our targets, which are mostly young (close to the ZAMS).

Although it is possible for δ Scuti pulsations to occur in the pre-main-sequence (PMS) phase, before the onset of hydrogen burning⁷⁴, there is no indication of a PMS classification in the literature for most of the stars in our sample.

Detailed modelling of HD 31901

As a member of the Pisces–Eridanus stellar stream, this star makes a good test case. We used the models described above, constrained by the observed frequencies of the radial and dipole modes and by the observed effective temperature and luminosity. Following Curtis et al.²⁶, we assumed the metallicity is close to solar. The results imply a mass of $(1.71 \pm 0.05)M_{\odot}$, a radius of $(1.54 \pm 0.03)R_{\odot}$ and an age of 150 ± 100 Myr. The latter is consistent with the age of about 130 Myr from Curtis et al.²⁶ but not with the value of about 1 Gyr determined by Meingast et al.²⁷.

Additional references and notes

As mentioned in the main text, several previous studies have reported regular frequency spacings in the Fourier amplitude spectra of δ Scuti stars^{14,18–20,48,75–86}. Among these, the following stars are included in our sample:

- HD 187547 (KIC 7548479): the large frequency spacing was previously reported as $40.5 \mu\text{Hz}$ (3.5 d^{-1})^{32,79}, which is a factor of two smaller than the value we have identified from the same Kepler observations. Comparing the échelle diagram of this star (Fig. 2) with others in our sample indicates that the larger $\Delta \nu$ is correct. This is also consistent with the Gaia DR2 parallax (6.57 ± 0.24 mas), which places this star close to the ZAMS.
- HD 34282 (V1366 Ori): based on observations with MOST (Microvariability and Oscillations of Stars), Casey et al.⁴⁸ reported a large separation of 3.75 d^{-1} , which is half the value reported here. Both values would be consistent with the HIPPARCOS parallax (5.24 ± 1.67 mas), as used by Casey et al., but the much more precise Gaia DR2 parallax (3.08 ± 0.29 mas) and comparison with other stars in our sample confirms that the larger $\Delta \nu$ value is correct. V1366 Ori is a Herbig Ae star⁸⁷, so it may be pre-main-sequence. Its classification in SIMBAD as an eclipsing binary appears to be incorrect.
- β Pictoris: known to be a high-frequency δ Scuti star^{88,89}, but a value for the large separation has not been reported. The TESS observations indicate a value of $\Delta \nu = 6.95 \text{ d}^{-1}$ (Fig. 4). The following stars are not in our sample but seem likely to be high-frequency δ Scuti stars with regular spacings:
- HD 144277: based on data from MOST and CoRoT (Convection, ROTation and planetary Transits), Zwintz et al.⁸⁰ suggested a large separation of 7.2 d^{-1} . This star will not be observed by TESS in its nominal two-year mission⁹⁰, but is scheduled to be observed in sector 39.
- HD 261711: based on MOST and CoRoT data, Zwintz et al.⁸¹ suggested a large separation of 6.72 d^{-1} . This star was observed by TESS in sector 6, but only with 30-min sampling.
- HD 174966: based on CoRoT data, García Hernández et al.⁸³ suggested a large separation of 5.53 d^{-1} . This star will not be observed by TESS in its nominal two-year mission⁹⁰.
- XX Pyx: based on ground-based multisite observations, Handler et al.⁷⁵ reported 22 pulsation frequencies in the range $27\text{--}76 \text{ d}^{-1}$ and suggested a large separation of 4.63 d^{-1} . We have examined the published frequencies for this star using échelle diagrams and confirm that a value of $\Delta \nu = 4.70 \text{ d}^{-1}$ gives a reasonably good alignment of the peaks. This star will not be observed by TESS in its nominal two-year mission⁹⁰, but is scheduled to be observed in sector 35.
- HD 156623: based on observations with the bRIng robotic observatory network, Mellon et al.⁹¹ found frequencies in the range $60\text{--}70 \text{ d}^{-1}$ and suggested regularity at three different separations: $3.75, 7.25$ and 2.75 d^{-1} . This star was observed by TESS in sector 12 and shows a pattern similar to other stars in our sample, with a spacing of $\Delta \nu = 7.31 \text{ d}^{-1}$.
- HD 27462 (TT Ret): based on TESS data, Khalack et al.⁹² preferred a large separation of 3.3 d^{-1} . Our examination of the TESS data and a comparison with the stars in our sample suggests $\Delta \nu = 6.9 \text{ d}^{-1}$. The WDS catalogue⁹³ lists this star as a binary with a separation of 0.4 arcsec and a magnitude difference of 0.7 . This is consistent with Gaia DR2, which gives no parallax and a large astrometric excess noise

($\text{RUWE} \approx 77$). Accounting for the binary, the HIPPARCOS parallax places the two components close to the ZAMS, consistent with our suggested value of Δv .

Data availability

TESS and Kepler data are available from the MAST portal (<https://archive.stsci.edu/access-mast-data>). All other data are available from the corresponding author upon reasonable request.

Code availability

We have made use of standard data analysis tools in Python, as noted and referenced in Methods.

33. Ricker, G. R. et al. Transiting Exoplanet Survey Satellite (TESS). *J. Astron. Telesc. Instrum. Syst.* **1**, 014003 (2014).
34. Borucki, W. J. et al. Kepler planet-detection mission: introduction and first results. *Science* **327**, 977 (2010).
35. MAST: Barbara A. Mikulski Archive for Space Telescopes (Space Telescope Science Institute, 2019); <https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html>.
36. Hey, D. & Ball, W. *Echelle: Dynamic Echelle Diagrams for Asteroseismology* v.1.4 (2020); <https://doi.org/10.5281/zenodo.3629933>.
37. Høg, E. et al. The Tycho-2 catalogue of the 2.5 million brightest stars. *Astron. Astrophys.* **355**, L27–L30 (2000).
38. Bessell, M. S. The Hipparcos and Tycho photometric system passbands. *Publ. Astron. Soc. Pacif.* **112**, 961–965 (2000).
39. Flower, P. J. Transformations from theoretical Hertzsprung–Russell diagrams to color-magnitude diagrams: effective temperatures, B–V colors, and bolometric corrections. *Astrophys. J.* **469**, 355–365 (1996).
40. Lindegren, L. et al. Gaia Data Release 2. The astrometric solution. *Astron. Astrophys.* **616**, A2 (2018).
41. Bovy, J., Rix, H.-W., Green, G. M., Schlafly, E. F. & Finkbeiner, D. P. On Galactic density modeling in the presence of dust extinction. *Astrophys. J.* **818**, 130 (2016).
42. Huber, D. et al. The K2 Ecliptic Plane Input Catalog (EPIC) and stellar classifications of 138,600 targets in campaigns 1–8. *Astrophys. J.* **224**, 2 (2016).
43. Choi, J. et al. Mesa Isochrones and Stellar Tracks (MIST). I. Solar-scaled models. *Astrophys. J.* **823**, 102 (2016).
44. Casagrande, L. et al. New constraints on the chemical evolution of the solar neighbourhood and Galactic disc(s). Improved astrophysical parameters for the Geneva–Copenhagen Survey. *Astron. Astrophys.* **530**, A138 (2011).
45. Huber, D. et al. Revised stellar properties of Kepler targets for the quarter 1–16 transit detection run. *Astrophys. J. Suppl. Ser.* **211**, 2 (2014).
46. Mathur, S. et al. Revised stellar properties of Kepler targets for the Q1–17 (DR25) transit detection run. *Astrophys. J. Suppl. Ser.* **229**, 30 (2017); erratum **234**, 43 (2018).
47. Brown, T. M., Latham, D. W., Everett, M. E. & Esquerdo, G. A. Kepler input catalog: photometric calibration and stellar classification. *Astron. J.* **142**, 112 (2011).
48. Casey, M. P. et al. MOST observations of the Herbig Ae δ -Scuti star HD 34282. *Mon. Not. R. Astron. Soc.* **428**, 2596–2604 (2013).
49. Vogt, S. S. et al. HIRES: the high-resolution echelle spectrometer on the Keck 10-m Telescope. *Proc. SPIE* **2198**, 362 (1994).
50. Howard, A. W. et al. The California Planet Survey. I. Four new giant exoplanets. *Astrophys. J.* **721**, 1467–1481 (2010).
51. Siverd, R. J. et al. NRES: the network of robotic echelle spectrographs. *Proc. SPIE* **10702**, 107026C (2018).
52. Brown, T. M. et al. Las Cumbres Observatory global telescope network. *Publ. Astron. Soc. Pacif.* **125**, 1031 (2013).
53. Gilbert, J. et al. Veloce Rosso: Australia's new precision radial velocity spectrograph. *Proc. SPIE* **10702**, 107020Y (2018).
54. Smith, K. C. & Dworetzky, M. M. In *Elemental Abundance Analyses* (eds Adelman, S. J. & Lanz, T.) 32–37 (Institut d'Astronomie de l'Université de Lausanne, 1988).
55. Smith, K. C. The Chemical Compositions of Mercury–Manganese Stars from Ultraviolet Spectra. PhD thesis, Univ. London (1992).
56. Castelli, F., Gratton, R. G. & Kurucz, R. L. Notes on the convection in the ATLAS9 model atmospheres. *Astron. Astrophys.* **318**, 841–869 (1997).
57. Kupka, F., Piskunov, N., Ryabchikova, T. A., Stempels, H. C. & Weiss, W. W. VALD-2: Progress of the Vienna atomic line data base. *Astron. Astrophys. Suppl. Ser.* **138**, 119–133 (1999).
58. Niemczura, E. et al. Spectroscopic survey of Kepler stars. I. HERMES/Mercator observations of A- and F-type stars. *Mon. Not. R. Astron. Soc.* **450**, 2764–2783 (2015).
59. Niemczura, E. et al. Spectroscopic survey of Kepler stars. II. FIES/NOT observations of A- and F-type stars. *Mon. Not. R. Astron. Soc.* **470**, 2870–2889 (2017).
60. Tkachenko, A. Grid search in stellar parameters: a software for spectrum analysis of single stars and binary systems. *Astron. Astrophys.* **581**, A129 (2015).
61. Tsymbal, V. STARS: A software system for the analysis of the spectra of normal stars. In *M.A.S.S., Model Atmospheres and Spectrum Synthesis* (eds Adelman, S. J., Kupka, F. & Weiss, W. W.) 198–199 (Astron. Soc. Pacif. Conf. Ser. Vol. 108, Astronomical Society of the Pacific, 1996).
62. Shulyak, D., Tsymbal, V., Ryabchikova, T., Stütz, C. & Weiss, W. W. Line-by-line opacity stellar model atmospheres. *Astron. Astrophys.* **428**, 993–1000 (2004).
63. Burgh, E. B. et al. Prime focus imaging spectrograph for the Southern African Large Telescope: optical design. *Proc. SPIE* **4841**, 1463–1471 (2003).
64. Kobulnicky, H. A. et al. Prime focus imaging spectrograph for the Southern African Large Telescope: operational modes. *Proc. SPIE* **4841**, 1634–1644 (2003).
65. Buckley, D. A. H., Swart, G. P. & Meiring, J. G. Completion and commissioning of the Southern African Large Telescope. *Proc. SPIE* **6267**, 62670Z (2006).
66. Kanodia, S. & Wright, J. Python leap second management and implementation of precise barycentric correction (barycorrpy). *Res. Not. AAS* **2**, 4 (2018).
67. Wright, J. T. & Eastman, J. D. Barycentric corrections at 1 cm s^{−1} for precise Doppler velocities. *Publ. Astron. Soc. Pacif.* **126**, 838–852 (2014).
68. Gagné, J. et al. BANYAN. XI. The BANYAN Σ multivariate Bayesian algorithm to identify members of young associations with 150 pc. *Astrophys. J.* **856**, 23 (2018).
69. Paxton, B. et al. Modules for Experiments in Stellar Astrophysics (MESA). *Astrophys. J. Suppl. Ser.* **192**, 3 (2011).
70. Paxton, B. et al. Modules for Experiments in Stellar Astrophysics (MESA): planets, oscillations, rotation, and massive stars. *Astrophys. J. Suppl. Ser.* **208**, 4 (2013).
71. Paxton, B. et al. Modules for Experiments in Stellar Astrophysics (MESA): binaries, pulsations, and explosions. *Astrophys. J. Suppl. Ser.* **220**, 15 (2015); erratum **223**, 18 (2016).
72. Ball, W. H. & Gizon, L. A new correction of stellar oscillation frequencies for near-surface effects. *Astron. Astrophys.* **568**, A123 (2014).
73. Asplund, M., Grevesse, N., Sauval, A. J. & Scott, P. The chemical composition of the Sun. *Annu. Rev. Astron. Astrophys.* **47**, 481–522 (2009).
74. Zwintz, K. et al. Echography of young stars reveals their evolution. *Science* **345**, 550–553 (2014).
75. Handler, G. et al. Delta Scuti Network observations of XX Pyx: detection of 22 pulsation modes and of short-term amplitude and frequency variations. *Mon. Not. R. Astron. Soc.* **318**, 511–525 (2000).
76. García Hernández, A. et al. Asteroseismic analysis of the CoRoT δ Scuti star HD 174936. *Astron. Astrophys.* **506**, 79–83 (2009).
77. Breger, M., Lenz, P. & Pamyatnykh, A. A. Towards mode selection in δ Scuti stars: regularities in observed and theoretical frequency spectra. *Mon. Not. R. Astron. Soc.* **396**, 291–298 (2009).
78. Breger, M. et al. Regularities in frequency spacings of δ Scuti stars: the Kepler star KIC 9700322. *Mon. Not. R. Astron. Soc.* **414**, 1721–1731 (2011).
79. Antoci, V. et al. The excitation of solar-like oscillations in a δ Sct star by efficient envelope convection. *Nature* **477**, 570–573 (2011).
80. Zwintz, K. et al. Regular frequency patterns in the classical δ Scuti star HD 144277 observed by the MOST satellite. *Astron. Astrophys.* **533**, A133 (2011).
81. Zwintz, K. et al. Regular frequency patterns in the young δ Scuti star HD 261711 observed by the CoRoT and MOST satellites. *Astron. Astrophys.* **552**, A68 (2013).
82. Paparo, M. et al. CoRoT 102749568: mode identification in a δ Scuti star based on regular spacings. *Astron. Astrophys.* **557**, A27 (2013).
83. García Hernández, A. et al. An in-depth study of HD 174966 with CoRoT photometry and HARPS spectroscopy. Large separation as a new observable for δ Scuti stars. *Astron. Astrophys.* **559**, A63 (2013).
84. Maceroni, C. et al. KIC 3858884: a hybrid δ Scuti pulsator in a highly eccentric eclipsing binary. *Astron. Astrophys.* **563**, A59 (2014).
85. Paparo, M., Benkő, J. M., Hareter, M. & Guzik, J. A. Unexpected series of regular frequency spacing of δ Scuti stars in the non-asymptotic regime. I. The methodology. *Astrophys. J.* **822**, 100 (2016).
86. Michel, E. et al. What CoRoT tells us about δ Scuti stars. Existence of a regular pattern and seismic indices to characterize stars. *Eur. Phys. J. Web Conf.* **160**, 03001 (2017).
87. Mora, A. et al. EXPORT: Spectral classification and projected rotational velocities of Vega-type and pre-main sequence stars. *Astron. Astrophys.* **378**, 116–131 (2001).
88. Mékarnia, D. et al. The δ Scuti pulsations of β Pictoris as observed by ASTEP from Antarctica. *Astron. Astrophys.* **608**, L6 (2017).
89. Zwintz, K. et al. Revisiting the pulsational characteristics of the exoplanet host star β Pictoris. *Astron. Astrophys.* **627**, A28 (2019).
90. Web TESS viewing tool (WTV) (TESS Science Support Center, 2020); <https://heasarc.gsfc.nasa.gov/cgi-bin/teess/webteess/wtv.py>.
91. Mellon, S. N. et al. Bright southern variable stars in the bRing survey. *Astrophys. J. Suppl. Ser.* **244**, 15 (2019).
92. Khalack, V. et al. Rotational and pulsational variability in the TESS light curve of HD 27463. *Mon. Not. R. Astron. Soc.* **490**, 2102–2111 (2019).
93. Mason, B. D., Wycoff, G. L., Hartkopf, W. I., Douglass, G. G. & Worley, C. E. The 2001 US Naval Observatory double star CD-ROM. I. The Washington double star catalog. *Astron. J.* **122**, 3466–3471 (2001).
94. Holdsworth, D. L. et al. High-frequency A-type pulsators discovered using SuperWASP. *Mon. Not. R. Astron. Soc.* **439**, 2078–2095 (2014).
95. Rodríguez, E., López-González, M. J. & López de Coca, P. A revised catalogue of δ Sct stars. *Astron. Astrophys. Suppl. Ser.* **144**, 469–474 (2000).
96. Amado, P. J. et al. The pre-main-sequence star HD34282: a very short-period δ Scuti-type pulsator. *Mon. Not. R. Astron. Soc.* **352**, L11–L15 (2004).
97. Gray, R. O. et al. The discovery of λ Bootis stars: the southern survey I. *Astron. J.* **154**, 31 (2017).
98. Murphy, S. J. et al. An evaluation of the membership probability of 212 λ Boo stars. I. A catalogue. *Publ. Astron. Soc. Aust.* **32**, e036 (2015).
99. Zuckerman, B., Rhee, J. H., Song, I. & Bessell, M. S. The Tucana/Horologium, Columba, AB Doradus, and Argus associations: new members and dusty debris disks. *Astrophys. J.* **732**, 61 (2011).
100. Torres, C. A. O. et al. Search for associations containing young stars (SACY). I. Sample and searching method. *Astron. Astrophys.* **460**, 695–708 (2006).
101. Murphy, S. J. & Lawson, W. A. New low-mass members of the Octans stellar association and an updated 30–40 Myr lithium age. *Mon. Not. R. Astron. Soc.* **447**, 1267–1281 (2015).
102. Paunzen, E. et al. λ Bootis stars in the SuperWASP survey. *Mon. Not. R. Astron. Soc.* **453**, 1241–1248 (2015).
103. Royer, F., Zorec, J. & Gómez, A. E. Rotational velocities of A-type stars. III. Velocity distributions. *Astron. Astrophys.* **463**, 671–682 (2007).

104. Royer, F., Grenier, S., Baylac, M. O., Gómez, A. E. & Zorec, J. Rotational velocities of A-type stars in the northern hemisphere. II. Measurement of $v \sin i$. *Astron. Astrophys.* **393**, 897–911 (2002).
105. Schröder, C., Reiners, A. & Schmitt, J. H. M. M. Ca II HK emission in rapidly rotating stars. Evidence for an onset of the solar-type dynamo. *Astron. Astrophys.* **493**, 1099–1107 (2009).

Acknowledgements We gratefully acknowledge the TESS and Kepler teams, whose efforts made these results possible. This research was partially conducted during the Exostar19 programme at the Kavli Institute for Theoretical Physics at UC Santa Barbara, which was supported in part by the National Science Foundation under grant no. NSF PHY-1748958. We thank colleagues in that programme, especially R. Townsend, for many stimulating discussions. We also thank A. Moya, A. G. Hernández, J. C. Suárez and Z. Guo for comments on the manuscript. We gratefully acknowledge support from the Australian Research Council (grant DE 180101104), and from the Danish National Research Foundation (grant DNRF106) through its funding for the Stellar Astrophysics Center (SAC). D.H. acknowledges support from the Alfred P. Sloan Foundation, the National Aeronautics and Space Administration (80NSSC18K1585, 80NSSC19K0379), and the National Science Foundation (AST-1717000). H.K. acknowledges support from the European Social Fund via the Lithuanian Science Council (LMTLT) grant 09.3.3-LMT-K-712-01-0103. Y.L. acknowledges support from the Joint Research Fund in Astronomy (U1631236) under cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS). D.L.H. acknowledges support by the Science and Technology Facilities Council under grant ST/M000877/1. The research leading to these results has (partially) received funding from the Research Foundation Flanders (FWO) under grant agreement G0H5416N (ERC Runner Up Project). This work makes use of observations from the LCOGT network. This work has also made use of data from the European Space Agency (ESA) mission Gaia (<https://www.cosmos.esa.int/gaia>), processed by the Gaia Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/>

consortium). Some of the observations reported in this paper were obtained with the Southern African Large Telescope (SALT) under programmes 2015-2-SCI-007, 2016-2-SCI-015 and 2017-2-SCI-010. The ISIS instrument is mounted on the WHT, which is operated on the island of La Palma by the Isaac Newton Group of Telescopes in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias. The Veloce Rosso facility was funded by Australian Research Council (ARC) Linkage Infrastructure, Equipment and Facility (LIEF) grants LE150100087 and LE160100014, and UNSW Research Infrastructure Scheme grant RG163088. C.G.T. and C.B. acknowledge the support of ARC Discovery grant DP170103491. V.A. was supported by a research grant (00028173) from VILLUM FONDEN. The authors wish to recognize and acknowledge the very significant cultural role and reverence that the summit of Mauna Kea has always had within the indigenous Hawaiian community; we are most fortunate to have the opportunity to conduct observations from this mountain. We also acknowledge the traditional owners of the land on which the Anglo-Australian Telescope stands, the Gamilaraay people, and pay our respects to elders past, present and emerging.

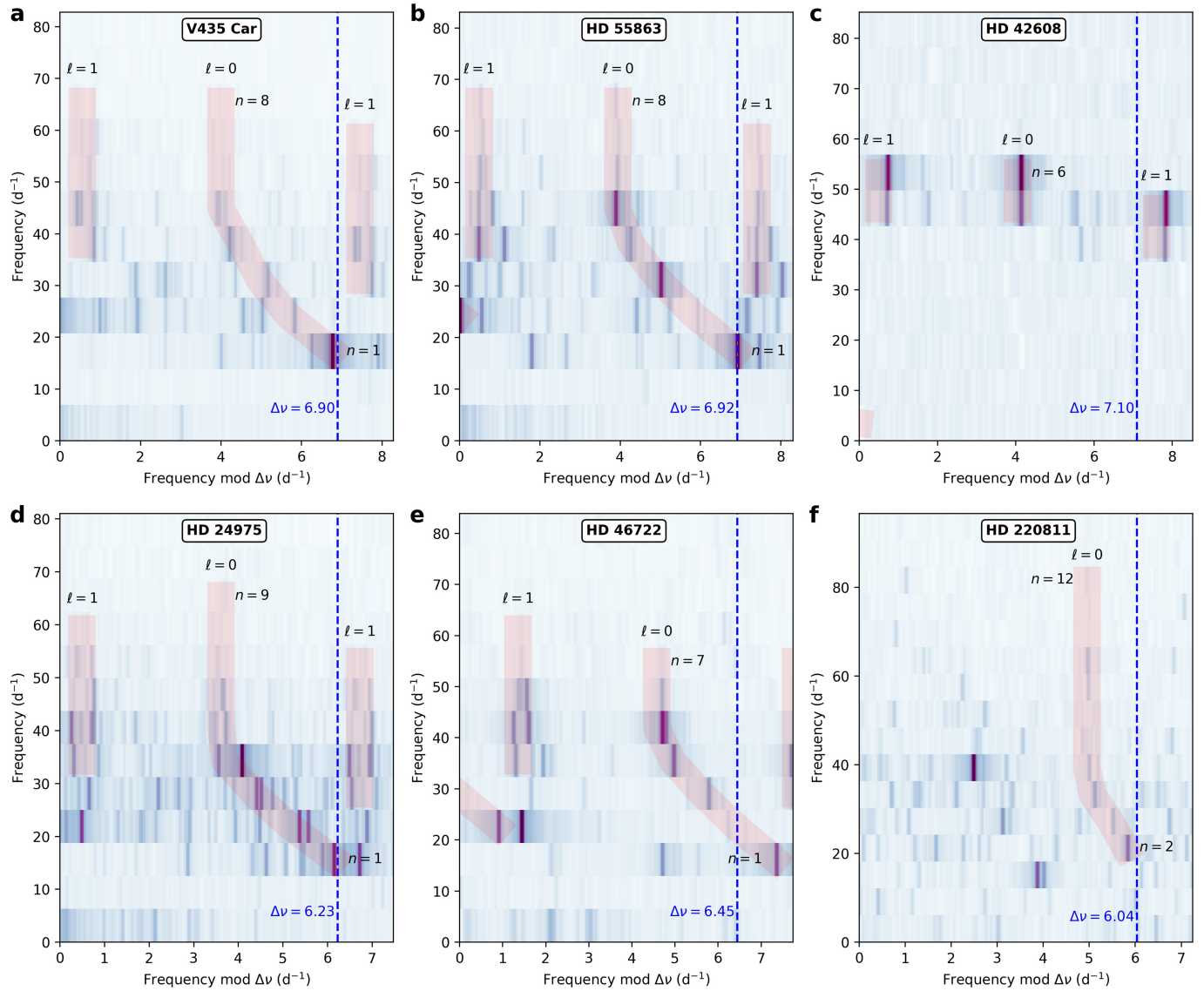
Author contributions T.R.B., S.J.M., D. R. Hey, W.J.C., G.L., Y.L., I.L.C. and J.Y. analysed the photometric observations; T.L., D.S., W.H.B., T.R.W., D.R.R., J.F. and J.J.H. calculated and/or interpreted theoretical models; V.A. and H.K. coordinated the selection of the targets for the TESS observations; D.H., D. R. Harbeck, S.S., B.S., T.M.B., A.W.H., H.I., C.M., M.R., C.B., A.D.R., C.G.T, M.J.I. and D.L.H. obtained and/or analysed the spectroscopic observations; E.G. and A.W.M. identified objects that belong to moving groups; G.R.R., R.K.V. and J.M.J. were key architects of the TESS Mission. All authors reviewed the manuscript.

Competing interests The authors declare no competing interests.

Additional information

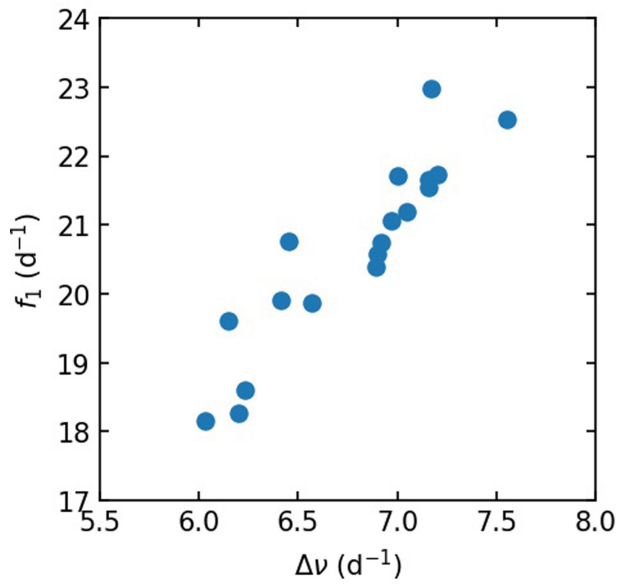
Correspondence and requests for materials should be addressed to T.R.B.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

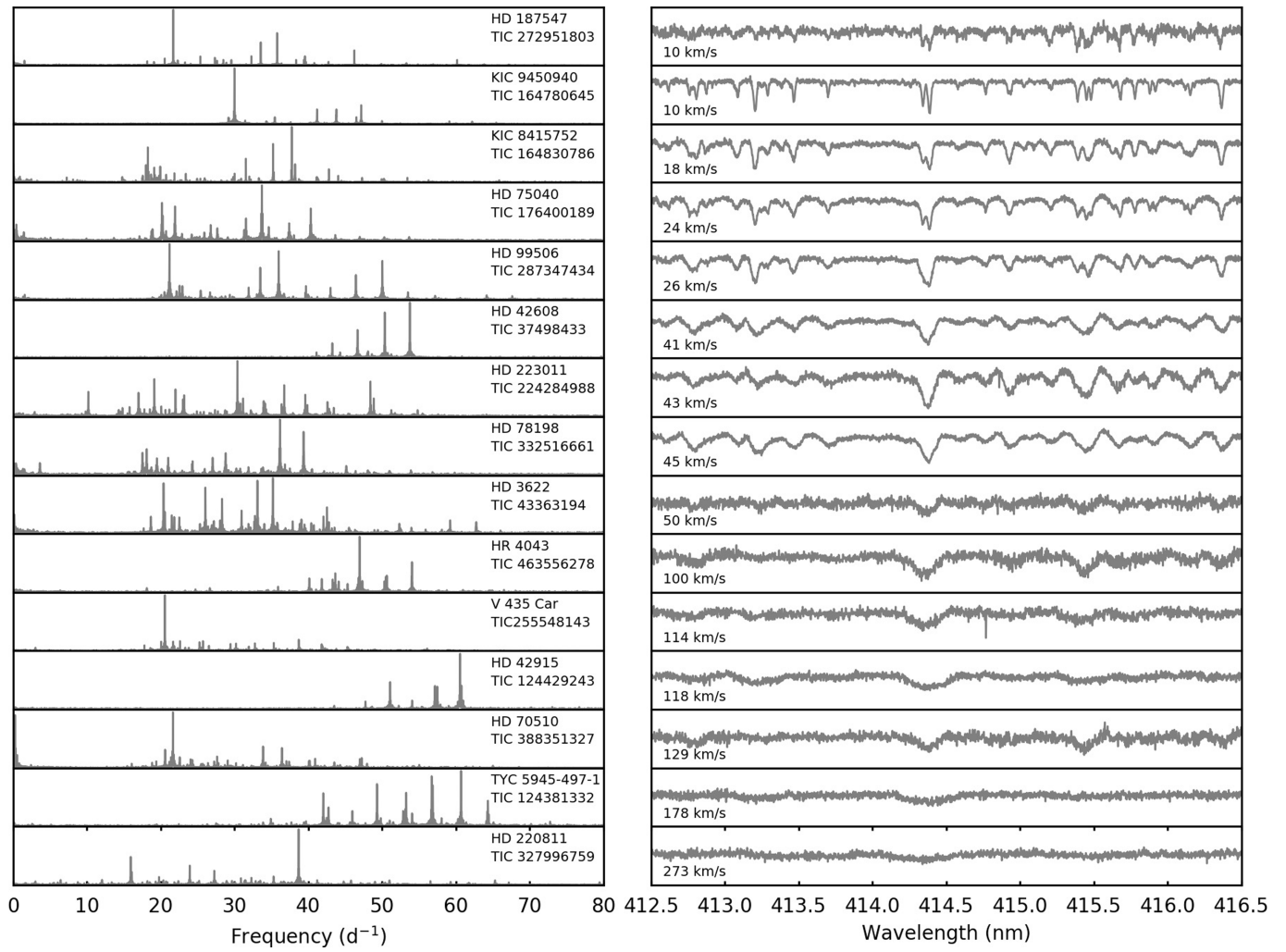


Extended Data Fig. 1 | More examples of mode identifications in δ Scuti stars. The amplitude spectra are shown in échelle format, segments of equal length being stacked vertically. **a**, V435 Car; **b**, HD 55863; **c**, HD 42608; **d**, HD 24975; **e**, HD 46722; **f**, HD 220811. The vertical dashed line shows the value of $\Delta\nu$ used in each case, with a repeated overlap region added on the right for

clarity. The greyscale shows the observed amplitude spectrum of data from the TESS spacecraft, where the number of 27-day sectors was four for V435 Car, three for HD 55863, two for HD 24975 and HD 46722, and one for HD 42608 and HD 220811. Smoothing was applied to the observed amplitude spectra before plotting, and the red stripes mark overtone sequences of $l=0$ and $l=1$ modes.



Extended Data Fig. 2 | Correlation between $\Delta\nu$ and the frequency of the fundamental radial mode. Symbols show 18 δ Scuti stars in which the fundamental radial mode, f_1 , is clearly identified. A correlation is expected because both quantities depend on the mean stellar density. We do not expect a perfect correlation owing to departures from the asymptotic relation¹⁻³ (see Methods) and variations in ε from star to star (see Fig. 3c).



Extended Data Fig. 3 | Fourier amplitude spectra and high-resolution spectra of high-frequency δ Scuti stars. Left panel, Fourier amplitude spectra of 15 stars; each star has two catalogue names, as shown. Right panel,

high-resolution spectra of the stars; measured $v \sin i$ values are given. The stars are sorted by increasing $v \sin i$ from top to bottom. See Methods section ‘High-resolution spectroscopy’ for details.

Extended Data Table 1 | Properties of high-frequency δ Scuti stars

HD	Name	TIC	<i>V</i>	<i>T</i> _{eff} (K)	<i>L</i> (<i>L</i> _⊙)	ρ (ρ_{\odot})	$\Delta\nu$ (d ⁻¹)	ε	Refs.
2280		281499618	9.13	7510	5.52 ± 0.26	0.49 ± 0.06	7.17	1.73	
3622		43363194	7.77	7930	7.86 ± 0.35	0.45 ± 0.06	6.89	1.61	
10779		229139161	8.78	7730	8.13 ± 0.36	0.39 ± 0.05	6.80	1.63	
10961		231014033	9.39	—	—	—	7.30	1.70	
17341		122615966	9.32	7810	10.05 ± 0.50	0.32 ± 0.05	5.90	1.73	97
		122686610	7.8	7880	10.21 ± 0.44	0.33 ± 0.04	6.41	1.61	
20203		274038922	8.85	7970	8.06 ± 0.38	0.45 ± 0.05	7.20	1.76	94
20232		159895674	6.88	8060	8.64 ± 0.36	0.44 ± 0.05	6.86	1.64	
24572		242944780	9.45	7410	7.25 ± 0.36	0.35 ± 0.05	7.20	1.58	94
24975		44645679	7.24	7790	9.20 ± 0.39	0.35 ± 0.04	6.23	1.58	
25248		459942890	8.6	—	—	—	7.16	1.55	
25369		9147509	9.68	—	—	—	6.15	1.65	
25674		34197596	8.69	8260	10.20 ± 0.50	0.42 ± 0.05	6.75	1.79	94
28548		71134596	9.22	8510	10.82 ± 0.55	0.45 ± 0.06	7.56	1.67	94,97
29783		269792989	7.87	—	—	—	6.74	1.79	
30422	EX Eri	589826	6.18	7940	8.42 ± 0.35	0.42 ± 0.05	6.52	1.86	95,98
31322		246902545	9.28	8260	13.19 ± 0.67	0.32 ± 0.04	6.10	1.69	94
31640		259675399	8.06	7690	8.25 ± 0.35	0.37 ± 0.05	6.41	1.75	
31901		316920092	9.07	7770	7.74 ± 0.39	0.41 ± 0.05	6.97	1.56	27
32433		348792358	9.22	7700	7.32 ± 0.35	0.42 ± 0.05	6.95	1.54	
34282	V1366 Ori	24344701	9.92	—	—	—	7.40	1.71	48,96
37286	HR 1915	31475829	6.26	8080	8.18 ± 0.34	0.47 ± 0.06	7.30	—	99
38597		100531058	8.65	8430	10.38 ± 0.47	0.44 ± 0.05	6.90	1.68	94
38629		32763133	8.92	8170	11.27 ± 0.53	0.35 ± 0.04	7.35	1.77	94
39060	β Pic	270577175	3.85	8080	8.49 ± 0.39	0.45 ± 0.05	6.95	—	25,88,89
40317		282265535	8.45	8700	10.58 ± 0.55	0.51 ± 0.06	6.95	1.63	
42005		408906554	9.54	8030	8.75 ± 0.42	0.42 ± 0.05	7.16	1.57	94
42608		37498433	9.85	8170	10.05 ± 0.49	0.40 ± 0.05	7.10	1.57	94
42915		124429243	9.04	8520	12.82 ± 0.68	0.38 ± 0.05	6.40	—	94,100,101
44726		150272131	10.38	7890	7.87 ± 0.38	0.44 ± 0.05	7.25	1.70	94
44930		34737955	9.42	7320	7.17 ± 0.40	0.33 ± 0.05	6.03	1.67	97
44958	V435 Car	255548143	6.74	7660	7.82 ± 0.32	0.38 ± 0.05	6.90	1.58	95
45424		117766204	7.18	8060	10.39 ± 0.44	0.36 ± 0.04	6.54	1.86	
46722		172193026	9.29	7810	8.28 ± 0.40	0.40 ± 0.05	6.45	1.71	97
48985		148228220	9.04	7710	11.60 ± 0.54	0.25 ± 0.03	7.25	1.56	
50153		78492107	7.03	7820	9.15 ± 0.39	0.36 ± 0.05	6.85	1.60	
54711		284348793	9.01	8200	9.22 ± 0.45	0.44 ± 0.06	7.08	1.65	
55863		294157254	9.06	7650	7.80 ± 0.38	0.38 ± 0.05	6.92	1.57	
59104		278179191	8.5	7360	6.15 ± 0.26	0.41 ± 0.05	6.96	1.67	
59594	V349 Pup	112484997	7.32	7800	8.06 ± 0.34	0.40 ± 0.05	6.65	1.61	95
67688		306773428	7.66	—	—	—	7.04	1.59	
70510		388351327	6.75	—	—	—	7.16	1.68	
75040		176400189	9.05	—	—	—	6.64	1.58	
78198		332516661	9.5	7340	7.79 ± 0.42	0.31 ± 0.04	5.90	1.65	
89263	HR 4043	463556278	6.22	—	—	—	7.02	1.71	
99506		287347434	8.36	7970	7.58 ± 0.37	0.48 ± 0.05	7.05	1.59	94
220811		327996759	6.91	—	—	—	6.04	1.82	
222496		316806320	9.48	—	—	—	5.63	1.71	
223011		224284988	6.32	7830	10.49 ± 0.44	0.31 ± 0.04	5.93	1.68	
290750		11199304	9.77	9170	19.14 ± 1.13	0.35 ± 0.05	5.96	—	
290799	V1790 Ori	11361473	10.67	8780	13.21 ± 0.98	0.42 ± 0.06	7.55	1.65	98,102
	SAO 150524	143381070	9.46	8030	7.88 ± 0.39	0.48 ± 0.06	7.63	—	
	SAO 249859	349645354	9.79	7070	7.99 ± 0.38	0.23 ± 0.04	5.34	1.69	
	TYC 85-867-1	431695696	9.63	7961	8.85 ± 0.57	0.40 ± 0.05	7.26	1.61	
	TYC 5945-497-1	124381332	9.69	8270	10.02 ± 0.53	0.42 ± 0.05	7.45	1.61	94
	TYC 8533-329-1	260161111	10.7	8370	9.33 ± 0.51	0.48 ± 0.06	7.27	—	94
	TYC 8564-537-1	340358522	10.59	7490	7.30 ± 0.36	0.37 ± 0.05	7.15	1.54	
187547		KIC 7548479	8.4	7470	6.74 ± 0.29	0.40 ± 0.05	7.00	1.61	32,79
	TYC 3132-1272-1	KIC 8415752	10.67	7780	9.83 ± 0.52	0.32 ± 0.05	6.20	1.63	
		KIC 9450940	12.68	7920	8.59 ± 0.58	0.41 ± 0.06	6.15	1.67	

HD and TIC indicate star catalogues; *V*, apparent magnitude; *T*_{eff}, effective temperature (fractional uncertainties are 2%); *L*, luminosity; ρ , mean density; $\Delta\nu$, large frequency separation; ε , phase term. Stars without values for *T*_{eff}, *L* and ρ are close binaries whose parameters cannot be reliably calculated (see Methods for details). References indicate classifications as δ Scuti stars^{32,48,79,88,89,94–96}, λ Boötis stars^{97,98}, and members of young moving groups, clusters or stellar streams^{25,27,99–102}.

Extended Data Table 2 | Projected rotational velocities from high-resolution spectroscopy

HD	Name	TIC	$v \sin i$ (km s^{-1})	Source
2280		281499618	26.4 ± 1.3	AAT+Veloce
3622		43363194	50 ± 6	LCO+NRES
10779		229139161	91 ± 5	AAT+Veloce
10961		231014033	(33 ± 3)	AAT+Veloce
17693		122686610	14 ± 1	AAT+Veloce
20203		274038922	40 ± 25	SALT+RSS
20232		159895674	37 ± 3	AAT+Veloce
24975		44645679	88 ± 4	AAT+Veloce
25674		34197596	160 ± 35	SALT+RSS
28548		71134596	200 ± 50	WHT+ISIS
30422	EX Eri	589826	128	literature ¹⁰³
31322		246902545	200 ± 50	SALT+RSS
31640		259675399	136 ± 4	AAT+Veloce
31901		316920092	33 ± 4	LCO+NRES
34282	V1366 Ori	24344701	129 ± 11	literature ⁸⁷
37286	HR 1915	31475829	70	literature ¹⁰⁴
38597		100531058	150 ± 40	SALT+RSS
38629		32763133	160 ± 40	SALT+RSS
39060	β Pic	270577175	122	literature ¹⁰⁵
42005		408906554	130 ± 30	SALT+RSS
42608		37498433	41 ± 1	Keck+HIRES
42915		124429243	118 ± 5	Keck+HIRES
44726		150272131	130 ± 40	SALT+RSS
44958	V435 Car	255548143	114 ± 11	LCO+NRES
48985		148228220	40 ± 4	AAT+Veloce
54711		284348793	50 ± 2	AAT+Veloce
55863		294157254	99 ± 5	AAT+Veloce
70510		388351327	94 ± 10	LCO+NRES
75040		176400189	24 ± 3	Keck+HIRES
78198		332516661	45 ± 1	Keck+HIRES
89263	HR 4043	463556278	100 ± 7	Keck+HIRES
99506		287347434	26 ± 2	Keck+HIRES
220811		327996759	261 ± 40	Keck+HIRES & LCO+NRES
223011		224284988	43 ± 2	LCO+NRES
	TYC 5945-497-1	124381332	178 ± 37	Keck+HIRES
	TYC 8533-329-1	260161111	100 ± 30	SALT+RSS
187547		KIC 7548479	10 ± 2	literature ⁷⁹
	TYC 3132-1272-1	KIC 8415752	18 ± 1	Keck+HIRES
		KIC 9450940	10 ± 1	Keck+HIRES

Values in parentheses of projected rotational velocity, $v \sin i$, indicate a close binary (see Methods), meaning the measurement may not be reliable. Details of sources (rightmost column) are given in Methods and refs. ^{79,87,103-105}.


Precise test of quantum electrodynamics and determination of fundamental constants with HD^+ ions

<https://doi.org/10.1038/s41586-020-2261-5>

Received: 18 November 2018

Accepted: 12 February 2020

Published online: 6 May 2020

 Check for updates

S. Alighanbari¹, G. S. Giri¹, F. L. Constantin^{1,2}, V. I. Korobov³ & S. Schiller^{1✉}

Bound three-body quantum systems are important for fundamental physics^{1,2} because they enable tests of quantum electrodynamics theory and provide access to the fundamental constants of atomic physics and to nuclear properties. Molecular hydrogen ions, the simplest molecules, are representative of this class³. The metastability of the vibration–rotation levels in their ground electronic states offers the potential for extremely high spectroscopic resolution. Consequently, these systems provide independent access to the Rydberg constant (R_∞), the ratios of the electron mass to the proton mass (m_e/m_p) and of the electron mass to the deuteron mass (m_e/m_d), the proton and deuteron nuclear radii, and high-level tests of quantum electrodynamics⁴. Conventional spectroscopy techniques for molecular ions^{5–14} have long been unable to provide precision competitive with that of *ab initio* theory, which has greatly improved in recent years¹⁵. Here we improve our rotational spectroscopy technique for a sympathetically cooled cluster of molecular ions stored in a linear radiofrequency trap¹⁶ by nearly two orders in accuracy. We measured a set of hyperfine components of the fundamental rotational transition. An evaluation resulted in the most accurate test of a quantum-three-body prediction so far, at the level of 5×10^{-11} , limited by the current uncertainties of the fundamental constants. We determined the value of the fundamental constants combinations $R_\infty m_e (m_p^{-1} + m_d^{-1})$ and m_p/m_e with a fractional uncertainty of 2×10^{-11} , in agreement with, but more precise than, current Committee on Data for Science and Technology values. These results also provide strong evidence of the correctness of previous key high-precision measurements and a more than 20-fold stronger bound for a hypothetical fifth force between a proton and a deuteron.

Since the inception of quantum mechanics, the precise understanding of three-body systems has represented a challenging fundamental physics problem. Its detailed study, both theoretical and experimental, is an ongoing effort, with a strong rate of improvement. Different three-body systems (for example, the helium atom, lithium ion, helium-like ions, antiprotonic helium atom and molecular hydrogen ions (MHIs)) provide the opportunity to test our understanding of quantum physics at the highest levels, in particular, the theory of quantum electrodynamics (QED). In doing so, important fundamental constants of physics (such as the Rydberg constant R_∞ , fine-structure constant α , electron mass m_e , proton mass m_p , deuteron mass m_d and antiproton mass) and particular nuclear properties, such as charge radii, electric quadrupole moments and charge-current moments, can be determined.

The MHIs (HD^+ , H_2^+ and so on) are molecular three-body systems containing two heavy particles and one light particle (electron). The electronic ground state supports hundreds of metastable rotation–vibration levels. A small subset of them have been studied with

different experimental techniques and concerning different aspects since the mid-1960s^{5–14,17} (for an early review, see ref. ³). Over the past decade, the MHIs have come into focus because of their relevance for the metrology of the particle masses^{4,18–21}. These can be determined from rotation–vibration spectroscopic data, an approach independent of the established technique of mass spectrometry in ion traps. An additional opportunity is the determination of the Rydberg constant R_∞ and the proton charge radius, independently from the established technique of atomic hydrogen spectroscopy^{22–24}. The precise value of these constants has been called into question in recent years in connection with the ‘proton radius puzzle’²⁵, and therefore alternative and independent approaches for its determination are highly desirable.

The *ab initio* theory of the MHIs has made enormous progress in precision over the past 20 years^{26–28}, reducing the uncertainty by four orders of magnitude. It currently stands at 1.4×10^{-11} fractionally for the fundamental rotational transition frequency and 7×10^{-12} for

¹Institut für Experimentalphysik, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany. ²Laboratoire PhLAM CNRS UMR 8523, Université Lille 1, Villeneuve d’Ascq, France. ³Bogoliubov Laboratory of Theoretical Physics, Joint Institute for Nuclear Research, Dubna, Russia. ✉e-mail: step.schiller@hhu.de

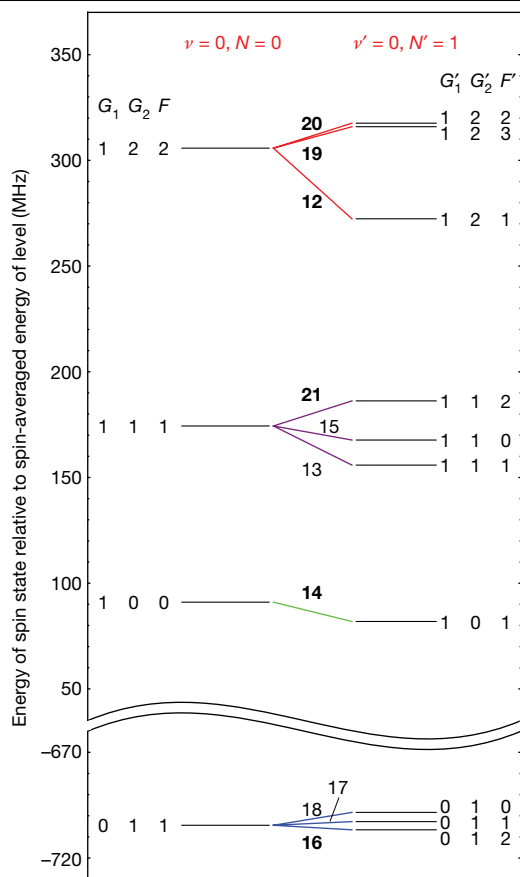


Fig. 1 | Energy diagram of the spin structures and favoured transitions.

The left side shows the rovibrational ground level ($v = 0, N = 0$) and the right side shows the rotationally excited level ($v' = 0, N' = 1$). The magnetic field is zero. The spin states are labelled by the (in part approximate) quantum numbers (G_1, G_2 and F). The spin energies $E_{\text{spin}}(v, N, G_1, G_2, F)$ and $E_{\text{spin}}(v', N', G'_1, G'_2, F')$ are shown as thin black lines. Transitions ('hyperfine components') are numbered according to increasing values of $hf_{\text{spin},i} = E_{\text{spin}}(v', N', G'_1, G'_2, F') - E_{\text{spin}}(v, N, G_1, G_2, F)$, including both favoured and weak transitions. The favoured electric-dipole transitions obey the selection rules $\Delta G_1 = 0, \Delta G_2 = 0$ and $\Delta F = 0, \pm 1$. The ten favoured transitions are shown by coloured lines. The rotational transition frequency of a particular hyperfine component is $f_i = f_{\text{spin-avg}} + f_{\text{spin},i}$, with $f_{\text{spin-avg}} \approx 1.314$ THz and, for favoured transitions, $f_{\text{spin},i} \approx \mathcal{O}(10$ MHz). The six components measured in this work are shown by bold numbers in the diagram.

vibrational transitions, ignoring spin-structure effects¹⁵. These uncertainties are smaller than the current (Committee on Data for Science and Technology (CODATA) 2018²⁹) uncertainties of the masses m_e, m_p and m_d , pointing at the potential of MHI spectroscopy for the metrology of fundamental constants. Here we perform precision spectroscopy of the fundamental rotational transition of HD^+ . Fundamental constants can be derived by comparison of the measured transition frequency $f^{(\text{exp})}$ with the prediction $f^{(\text{theor})} = 2cR_\infty(m_e/\mu_{\text{pd}})F_{\text{spin-avg}}^{(\text{theor})}$, where $\mu_{\text{pd}} = m_p m_d / (m_p + m_d)$ is the reduced nuclear mass, c is the speed of light, and $F_{\text{spin-avg}}^{(\text{theor})} = 0.244591781951(33)_{\text{theory}}(11)_{\text{CODATA2018}}$ is a dimensionless normalized frequency computed ab initio, neglecting the hyperfine interactions. $F_{\text{spin-avg}}^{(\text{theor})}$ encompasses—besides the dominant non-relativistic (Schrödinger) part—essential relativistic, nuclear-size-related and radiative contributions. The nuclear charge radius values (r_p, r_d) are from the CODATA 2018 adjustment that took into account the muonichydrogen spectroscopy results. Whereas the uncertainty of $F_{\text{spin-avg}}^{(\text{theor})}$ due to theory is 1.4×10^{-11} , the uncertainty originating from the CODATA 2018 uncertainties of the fundamental constants is smaller (4.4×10^{-12}), and stems from the uncertainties of r_p and r_d .

Apart from a matching comparison with a 50-year-old radiofrequency (RF) spectroscopy benchmark result on H_2^+ (ref.¹⁷), the ab initio theory could not be tested experimentally at a competitive level, owing to lack of suitable experimental methods. With a few exceptions, the spectroscopic resolution in rotational and vibrational spectroscopy of molecular ions in general has been limited by Doppler broadening. Although this broadening can be minimized by trapping molecular ions in an RF trap and sympathetically cooling them by atomic ions, their effective temperature remains of the order of 10 mK, leading to Doppler-limited linewidths not lower than 5×10^{-8} fractionally¹². Unresolved hyperfine structure increases linewidths again^{11,14}, posing a roadblock for testing theory at more precise levels.

Only recently, new methods have been introduced that open up the next generation of precision experiments^{30,31}. Specifically for rotational spectroscopy, we have shown¹⁶ that sub-Doppler spectroscopy is possible for a radiation propagation direction transverse to the 'long' axis of the molecular ion cluster (trapped ion cluster transverse excitation spectroscopy, TICTES). The small motional amplitude of the ions along the spectroscopy wave propagation direction compared with its wavelength allows reaching the Lamb–Dicke regime. In the first demonstration¹⁶, a fractional line resolution of 1×10^{-9} (full-width at half-maximum (FWHM) relative to absolute frequency) was obtained.

Here we improve the resolution of TICTES by more than two orders of magnitude. This enables a detailed direct study of the fundamental rotational transition of HD^+ , whose hyperfine spectrum and Zeeman splittings are resolved and systematic effects are determined.

Comparison with our improved theory and a new analysis method allows us to establish agreement between theory and experiment at the 5×10^{-11} level (limited by CODATA 2018 uncertainties), not only representing the most accurate test of a molecular three-body system so far, but also demonstrating the power of TICTES, a method applicable to a plethora of molecular ions.

The experiment

We performed spectroscopy of the fundamental rotational transition ($v, N = (0, 0) \rightarrow (v, N') = (0, 1)$) at 1.3 THz. v and N are the vibrational and rotational quantum numbers, respectively. See Extended Data Fig. 1 for the experimental scheme. The fractional population of HD^+ ions in the lower spectroscopy state (0, 0) is enhanced using rotational laser cooling³². The transition is detected by resonance-enhanced multiphoton dissociation (REMPD)³³. See Extended Data Fig. 2 for typical data. To achieve a spectroscopy wave with narrow linewidth, high frequency stability and high accuracy, a GPS-monitored, hydrogen-maser-referenced terahertz frequency multiplier is used^{16,34}. Compared with our previous work¹⁶, we performed measurements for different magnetic-, electric- and light-field strengths, and minimized the terahertz wave power. These extensive measurements were enabled by improvements in the long-term stability of the apparatus and improved detection schemes.

The HD^+ molecule has spin structure in both the lower and the upper rotational levels, due to the presence of (1) the intrinsic spins of the electron (s_e), proton (I_p) and deuteron (I_d), and (2) of the rotational angular momentum N (Fig. 1). For state description, we use the angular momentum coupling scheme $G_1 = s_e + I_p$, $G_2 = G_1 + I_d$, $F = G_2 + N$ (ref.³⁵), where F is the total angular momentum. The rotational transition encompasses 32 hyperfine components f_i in absence of a magnetic field; of these, ten are favoured (strong) (Fig. 1). Their frequencies f_{12}, \dots, f_{21} lie within a range of 45 MHz around $f_{\text{spin-avg}} \approx 1.314$ THz. Averaging over these ten components with appropriate weights yields the 'spin-averaged' frequency $f_{\text{spin-avg}}$ (ref.³⁶). Here we measured six hyperfine components, $f_{12}, f_{14}, f_{16}, f_{19}, f_{20}$ and f_{21} .

Figure 2 shows the measured transitions, in the presence of a small magnetic field. The different linewidths are due to the different terahertz wave intensities used and due to the different transition dipole moments. Line 19 includes the two transitions between states of

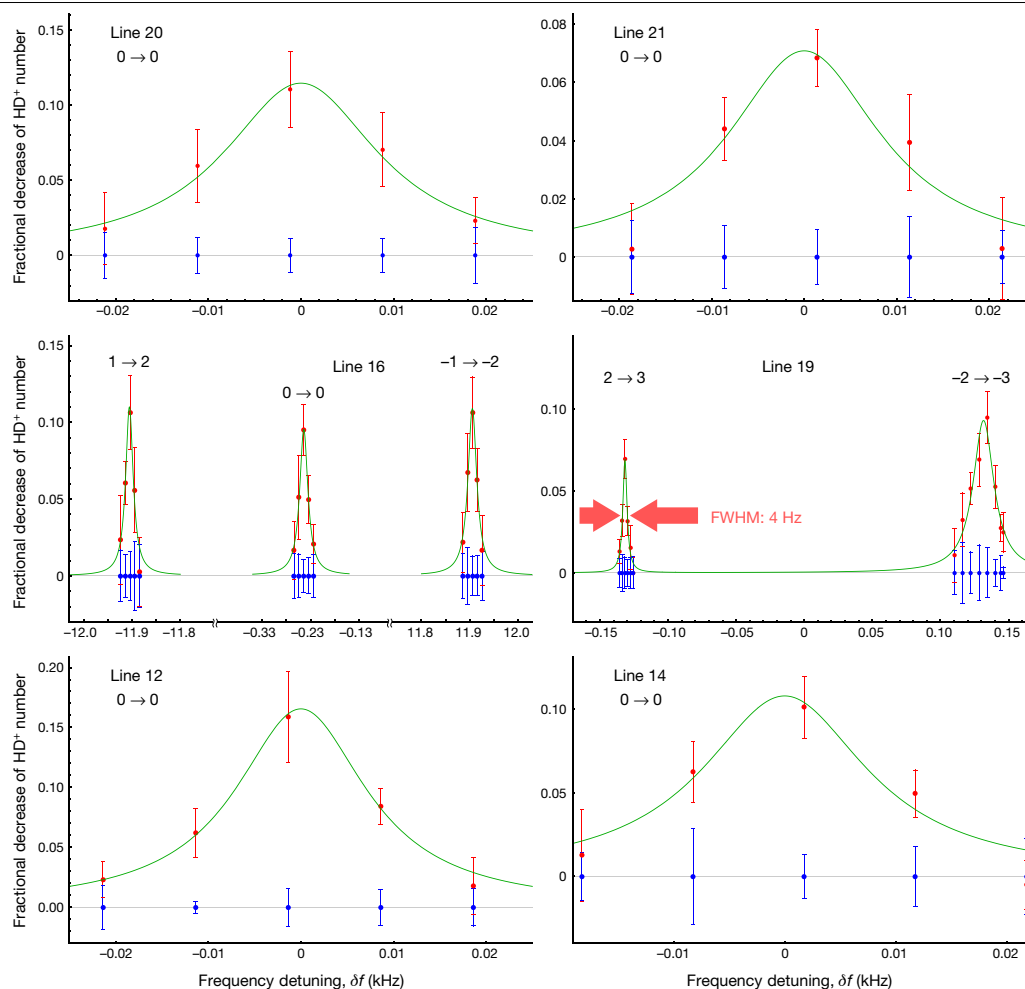


Fig. 2 | Hyperfine components of the fundamental rotational transition of HD^+ at 1.3 THz. The red and blue points indicate the cases of terahertz radiation on and off (background), respectively. Green lines are Lorentzian fits. The Zeeman components are indicated by the expression $m_f \rightarrow m'_f$. The terahertz wave intensity varied and was less than 10 nW mm^{-2} . The zero of the

frequency scales are set to coincide with the fitted line maxima or means. At each frequency setting, the red and blue data points are both shown with an offset equal to the value of the blue point. Each error bar represents the standard deviation of the mean. The nominal magnetic field is $B_{\text{nom}} \approx 30 \text{ } \mu\text{T}$ and the trap RF amplitude is approximately 190 V.

maximum total angular momentum F and maximum projection quantum number m_f , ($F=2, m_f=\pm 2$) \rightarrow ($F'=3, m'_f=\pm 3$), denoted by 19_{\pm} , whose Zeeman shift is purely linear, according to theory³⁷. The two components were observed at lower resolution and with unresolved Zeeman splitting in ref.¹⁶. One Zeeman component (19_{-}) measured at particularly low intensity exhibited a full linewidth of 4 Hz, or 3×10^{-12} fractionally, indicating the potential of the experimental technique in the context of mass determination. For line 16, we measured a Zeeman pair $m_f = \pm 1 \rightarrow m'_f = \pm 2$ (denoted by 16_{\pm}), split by a linear Zeeman shift and weakly shifted by a common quadratic Zeeman shift, and a component $16_0: m_f = 0 \rightarrow m'_f = 0$, which exhibits a moderate quadratic Zeeman shift³⁷. For the remaining lines, we measured only the $m_f = 0 \rightarrow m'_f = 0$ Zeeman components.

Systematic shifts

For an accurate comparison between theoretical transition frequencies (computed assuming an absence of perturbing fields) and experimental values (measured in presence of such fields), the systematic shifts must be taken into account. We determined them experimentally. The dominant systematic effect is the Zeeman shift. For a nominal RF drive amplitude, we measured the frequency shifts of all considered components as a function of applied magnetic field. The shifts are consistent with the theoretically calculated ones, except for small deviations.

We obtained the transition frequencies corresponding to zero magnetic field by extrapolation.

The quadratic Stark shift due to the ion trap's electric field $E(t)$, oscillating at comparatively low (RF) frequency and leading to a mean-square value $\langle E(t)^2 \rangle$, is a second shift, of lower magnitude. For a nominal magnetic field, we measured the frequency shifts of all considered components for a set of trap RF amplitudes. All shifts were found to increase with amplitude, with values in the range of 0.5 to 1.2 kHz kV^{-2} . We determined the frequencies corresponding to zero RF-field amplitude by extrapolation. For additional information, see Methods and Extended Data Fig. 3.

Table 1 presents the experimental transition frequencies $f_i^{(\text{exp})}$ (corrected for the systematic shifts) and their uncertainties. The uncertainties result from the number of frequency measurements, which were taken at different RF drive settings and different magnetic-field settings, and the statistical uncertainties of the frequency measurements. The lowest experimental uncertainty is achieved for line 16, $u(f_{16}^{(\text{exp})}) = 0.017 \text{ kHz}$ (fractional uncertainty $u_r = 1.3 \times 10^{-11}$). This represents the best performance level of the TICTES technique as currently implemented.

Theory

For a compelling comparison between theory and the experimental data, highly precise theoretical predictions and qualified estimates of

Table 1 | Experimental rotational frequencies, and comparison with theoretical ab initio frequencies

Line <i>i</i>	$G_i G_2 F \rightarrow G_i' G_2' F'$	$f_i^{(\text{exp})}$	$u(f_i^{(\text{exp})})$	$f_i^{(\text{theor})}$	$u(f_{\text{spin},i}^{(\text{theor})})$	$u(f_{\text{spin-avg}}^{(\text{theor})})$	$u_{\text{CODATA}}(f_i^{(\text{theor})})$
12	122 \rightarrow 121	1314892544.276	0.040	1314892544.23	1.2	0.018	0.061
14	100 \rightarrow 101	1314916678.487	0.064	1314916678.74	1.3	0.018	0.061
16	011 \rightarrow 012	1314923618.028	0.017	1314923617.94	0.20	0.018	0.061
19	122 \rightarrow 123	1314935827.695	0.037	1314935827.58	1.2	0.018	0.061
20	122 \rightarrow 122	1314937488.614	0.060	1314937488.80	1.4	0.018	0.061
21	111 \rightarrow 112	1314937540.762	0.046	1314937540.61	0.73	0.018	0.061

Uncertainties are denoted by u . Frequency values are in kHz. The theoretical values $f_i^{(\text{theor})}$ were computed using CODATA 2018 constants. The last three columns show the three contributions to the total uncertainty of $f_i^{(\text{theor})}$. Line 16 offers the most stringent comparison, due to its comparatively small theory uncertainty.

their uncertainties are essential. The ab initio transition frequency $f_i^{(\text{theor})}$ of each hyperfine component i is the sum of two contributions, $f_{\text{spin-avg}}^{(\text{theor})} + f_{\text{spin},i}^{(\text{theor})}$. The dominant contribution is

$$f_{\text{spin-avg}}^{(\text{theor})} = 1,314,925,752.896(18)_{\text{theory}}(61)_{\text{CODATA2018}} \text{ kHz} \quad (1)$$

computed¹⁵ including all relativistic and radiative corrections up to the relative order α^5 and partially including contributions of the order α^6 (Table 2). The value $f_{\text{spin-avg}}^{(\text{theor})}$ is updated from the value reported in ref.¹⁶ by using CODATA 2018²⁹ updates of the Rydberg constant, the particle masses (in atomic mass units, u), the proton charge radius and the deuteron charge radius. The theory uncertainty is estimated as $u(f_{\text{spin-avg}}^{(\text{theor})}) \approx 0.018$ kHz, while the larger CODATA 2018 uncertainty, $u_{\text{CODATA2018}}(f_{\text{spin-avg}}^{(\text{theor})}) \approx 0.061$ kHz, is dominated by the uncertainties of the particle masses.

A spin frequency contribution $f_{\text{spin},i}^{(\text{theor})}$ is the difference of the spin structure energies of the upper and lower spin states involved in the transition. For the favoured transitions measured here, the values of $f_{\text{spin},i}^{(\text{theor})}$ are of the order of 10 MHz. The spin contributions are computed by diagonalizing the Breit–Pauli spin Hamiltonian of ref.³⁵. The various terms of this Hamiltonian are proportional to coefficients $\mathcal{E}_k, \mathcal{E}_k'$, computed ab initio (Extended Data Table 1). The spin Hamiltonian of the $N=0$ level necessitates two coefficients, \mathcal{E}_4 and \mathcal{E}_5 , while the $N=1$ level necessitates nine, $\mathcal{E}_1', \dots, \mathcal{E}_9'$.

The coefficients $\mathcal{E}_4, \mathcal{E}_4'$ and $\mathcal{E}_5, \mathcal{E}_5'$ describe the dominant $\mathbf{s}_e \cdot \mathbf{I}_p$ and $\mathbf{s}_e \cdot \mathbf{I}_d$ interactions, respectively, and have been calculated with high theoretical precision, including all corrections of the order $\alpha^2 E_F/h$ and the leading corrections of the order $\alpha^3 E_F/h$, where $E_F \approx h(1.4 \text{ GHz})$ is the Fermi contact energy for the hyperfine splitting in atomic hydrogen and h is Planck's constant³⁸. The fractional theoretical uncertainties of these spin Hamiltonian coefficients are of the order α^3 ; they are estimated

as $\varepsilon_F = 1 \times 10^{-6}$. Furthermore, the signed theory errors are expected to be nearly equal: $\Delta \mathcal{E}_4^{(\text{theor})} \approx \Delta \mathcal{E}_4'^{(\text{theor})}$ and $\Delta \mathcal{E}_5^{(\text{theor})} \approx \Delta \mathcal{E}_5'^{(\text{theor})}$ (Methods).

The other spin coefficients, $\mathcal{E}_1', \mathcal{E}_2', \mathcal{E}_3', \mathcal{E}_6', \mathcal{E}_7', \mathcal{E}_8'$ and \mathcal{E}_9' , have been obtained within the Breit–Pauli approximation. We computed them using our most precise non-relativistic non-adiabatic molecular variational wave functions (Methods, Extended Data Table 1). The omitted terms are of the relative order α^2 . References^{38,39} lead us to estimate a common fractional theory uncertainty equal to $\alpha^2 = \varepsilon_0 \approx 5 \times 10^{-5}$.

To determine the impact of the theory uncertainty of a particular Hamiltonian coefficient on a particular spin frequency, we introduce the quantities $\gamma_{i,k}' \Delta \mathcal{E}_k^{(\text{theor})}$, with the derivatives $\gamma_{i,k}' = \partial E_{\text{spin},i}'(\mathcal{E}_1', \dots, \mathcal{E}_9') / \partial \mathcal{E}_k'$ relevant for the upper spin level and similarly for the lower spin level. The γ values are reported in Extended Data Table 1. Assuming equal theory errors for the pairs $(\mathcal{E}_4, \mathcal{E}_4')$ and $(\mathcal{E}_5, \mathcal{E}_5')$, we conservatively estimate the total theory uncertainty of the spin-frequency contribution with the following expression

$$u(f_{\text{spin},i}^{(\text{theor})}) = \varepsilon_F \sum_{4,5} |\gamma_{i,k}' \mathcal{E}_k' - \gamma_{i,k} \mathcal{E}_k| + \varepsilon_0 \sum_{1,2,3,6,7,8,9} |\gamma_{i,k}' \mathcal{E}_k'|$$

The form of the first sum embodies the assumption of equal fractional errors and correlation, $\Delta \mathcal{E}_{4,5}^{(\text{theor})} = \delta_4 \varepsilon_F \mathcal{E}_{4,5}$, $\Delta \mathcal{E}_{4,5}'^{(\text{theor})} = \delta_4 \varepsilon_F \mathcal{E}_{4,5}'$, with $\delta_4 = 1$ or -1 , $\delta_5 = 1$ or -1 . The similarities $\gamma_4 \approx \gamma_4'$ and $\gamma_5 \approx \gamma_5'$ for the lower and upper rotational levels then lead to a strong suppression of the contributions related to the theory errors of $\mathcal{E}_4, \mathcal{E}_4', \mathcal{E}_5$ and \mathcal{E}_5' . This results in the spin-frequency uncertainties shown in Table 1 (column 6). They dominate the total uncertainty of the transition frequencies $f_i^{(\text{theor})}$.

Comparison between theory and experiment

Table 1 presents the comparison between the theory and experimental data of the individual hyperfine components of the rotational

Table 2 | Contributions to the ab initio spin-averaged rotational frequency $f_{\text{spin-avg}}^{(\text{theor})}$

Term	Relative order	Contribution (kHz)	Origin
$f^{(0)}$	1	1,314,886,776.526	Solution of three-body Schrödinger equation
$f^{(2)}$	α^2	48,416.268	Relativistic corrections in Breit–Pauli approximation; nuclear radii
$f^{(3)}$	α^3	−9,378.119	Leading-order radiative corrections (for example, leading-order Lamb shift, anomalous magnetic moment)
$f^{(4)}$	α^4	−65.631(2)	One-loop, two-loop radiative corrections; relativistic corrections
$f^{(5)}$	α^5	3.923(3)	Radiative corrections up to three-loop diagrams; Wichman–Kroll contribution
$f^{(6)}$	α^6	−0.070(18)	Higher-order radiative corrections
Total $f_{\text{spin-avg}}^{(\text{theor})}$		1,314,925,752.896(18)	

The values were calculated using CODATA 2018 values of the fundamental constants. The main contribution $f^{(0)}$ is of order $cR_\infty(m_e/\mu_{\text{ad}})$. Recoil corrections (due to finite masses of nuclei) are included fully at the order α^2 ; the leading recoil corrections proportional to m_e/m_p or m_e/m_d are included at the order α^3 . Contributions due to the finite size of the nuclei are included in the $f^{(2)}$ term¹⁵. The one-loop contribution from $\mu^+ - \mu^-$ vacuum polarization is included in $f^{(3)}$. The estimated fractional theory uncertainty of the spin-averaged frequency is $u_i = 1.4 \times 10^{-11}$ ($u(f_{\text{spin-avg}}^{(\text{theor})}) = 0.018$ kHz). The impact of the fundamental constants' uncertainties is given in the text. The change in the value of $f^{(0)}$ from CODATA 2014 to CODATA 2018 has contributions of −0.041 kHz from the Rydberg constant adjustment and 0.213 kHz from the particle masses adjustments. The change in the value of $f^{(2)}$ due to the proton and deuteron charge radii adjustments is 0.104 kHz.

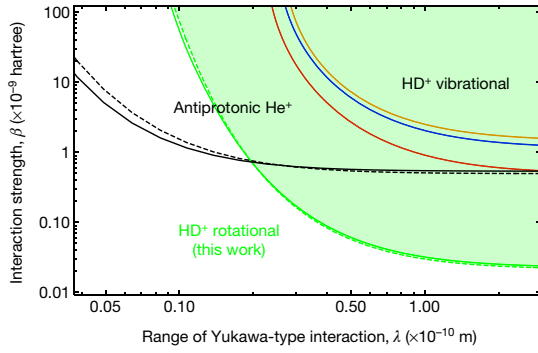


Fig. 3 | Exclusion plot (95% confidence limit) for a Yukawa-type interaction between a proton and a deuteron, deduced from spectroscopy of MHIs.

The parameter space above the lines is excluded. The assumed interaction is $V_5(R) = \beta N_1 N_2 \exp(-R/\lambda)/R$, where R is the proton–deuteron distance, λ is the interaction range, $N_1 = 1$ and $N_2 = 2$ are the nuclear mass numbers, and β is the interaction strength. Green lines, this work (full green, numerical; dashed green, analytical, equation (4) in Methods); red line, ref. ¹⁴; blue line, ref. ¹¹; orange line, ref. ¹². For comparison, the black lines show the limits for the interaction between the antiproton and the helium-4 nucleus, obtained from two different transitions⁴⁶. See Methods for details.

transition. We find agreement for all lines, within the combined uncertainties of theory and experiment. The agreement is most stringent for line 16, and it is limited by the prediction's total uncertainty $u(f_{16}^{(\text{theor})}) \approx 0.21$ kHz, or 1.5×10^{-10} fractionally. The agreement is far less stringent than the roughly ten times lower experimental uncertainty would allow. The precise experimental value can therefore serve as a benchmark for tests of future improved spin-structure calculations.

Frequencies related to only the spin structure of the molecule can be obtained from rotational frequency differences $\Delta f_{ij} = f_i - f_j = f_{\text{spin},i} - f_{\text{spin},j}$, where the spin-averaged frequency is cancelled. All deviations between experiment and theory are smaller than 0.42 kHz in magnitude and are well within the theory uncertainties (CODATA 2018 uncertainties are not relevant here). The most stringent theory–experiment agreement is found for $\Delta f_{21,19}$, within the roughly 0.7-kHz theory uncertainty, but ten times less stringent than the experimental uncertainty would allow.

In view of the relatively large uncertainties for $f_{\text{spin},i}^{(\text{theor})}$ above, we introduce a novel way of comparing experiment with theory, using composite frequencies defined as $f_c = \sum_i b_i f_i$, with appropriate weights b_i . We aim to find composite frequencies with small theory uncertainty, and therefore must suppress the contribution of the spin energies' uncertainties without suppressing the spin-averaged energies that give rise to $f_{\text{spin-avg}}$. The latter requirement is satisfied by imposing the 'normalization' condition $\sum_i b_i = 1$, so that $f_c = f_{\text{spin-avg}} + f_{\text{spin},c}$, with $f_{\text{spin},c} = \sum_i b_i f_{\text{spin},i}$. The former requirement is implemented by finding the composite frequency that minimizes the theory uncertainty. We use a conservative measure of theory uncertainty that does not assume any relationship between the theory errors of $(\mathcal{E}_i, \mathcal{E}'_i)$ and of $(\mathcal{E}_j, \mathcal{E}'_j)$: $\bar{u}(f_{\text{spin},c}^{(\text{theor})}) = \sum_k (|\sum_i b_i \gamma'_{i,k} \mathcal{E}'_k| + |\sum_i b_i \gamma_{i,k} \mathcal{E}_k|) \mathcal{E}_k$. The solution $\{b_i\}$ is found numerically (see 'Composite frequencies' in Methods). $f_{\text{spin},c}^{(\text{theor})}(\{b_i\}) = 934.635$ kHz, with negligible uncertainty $\bar{u}(f_{\text{spin},c}^{(\text{theor})}) = 0.001$ kHz. We note that this approach for eliminating the spin-energy-related uncertainty is complementary to the more general method recently proposed by some of us in ref. ³⁶, where the composite frequency is equal to $f_{\text{spin-avg}}$.

From the experimental composite frequency, we deduce the experimental spin-averaged frequency

$$f_{\text{spin-avg}}^{(\text{exp})} = f_c^{(\text{exp})}(\{b_i\}) - f_{\text{spin},c}^{(\text{theor})}(\{b_i\}) = 1,314,925,752.910(17)_{\text{exp}} \text{ kHz} \quad (2)$$

($u_r = 1.3 \times 10^{-11}$). The theory uncertainty (via $f_{\text{spin},c}^{(\text{theor})}$) is negligible and is therefore not indicated.

QED test and determination of fundamental constants

A comparison of equations (1) and (2) indicates that our experiment and theory achieve a successful test of three-body physics with a combined fractional uncertainty of 4.8×10^{-11} (0.064 kHz), limited by CODATA 2018 uncertainties. Comparing the total uncertainty of $f_{\text{spin-avg}}^{(\text{theor})}$ with the QED contributions listed in Table 2, we see that it is close to the QED contribution of highest calculated relative order, $f^{(6)} \approx 0.070(18)$ kHz. Therefore, more specifically, our experiment furnishes a test of QED at the relative order of α^6 . According to theory, the contributions to $f^{(2)}$ stemming from the finite proton root-mean-square charge radius r_p and the deuteron charge radius r_d with their CODATA 2018 uncertainties are $-0.644(3)$ kHz and $-4.120(3)$ kHz, respectively. The sum of these contributions is put in evidence by our experiment–theory comparison, with a fractional uncertainty of 1.4%.

Our experiment–theory agreement is obtained when including in the hyperfine structure calculation the contribution of the deuteron quadrupole moment Q_d , quantified by the coefficient $\mathcal{E}'_9 \propto Q_d$. This contribution is observed here in an MHI for the first time. From the measured hyperfine structure we can extract, independently of any QED contributions, a value for Q_d with 1.5% fractional uncertainty (Methods).

The experiment–theory agreement can also be used to set improved limits to the hypothetical existence of a spin-averaged fifth force between a proton and a deuteron (Fig. 3, Methods). Compared with previous bounds from MHI spectroscopy, the improvement is a factor of 21 or more for force ranges $\lambda > 1$ Å.

We can obtain the combination $R_\infty m_e / \mu_{\text{pd}}$ of fundamental constants from any of the measured rotational frequencies $f_i^{(\text{exp})}$ and the corresponding ab initio value $f_i^{(\text{theor})}$. However, the highest precision is obtained by instead choosing the composite frequency f_c or the spin-averaged frequency, because their spin-structure theory uncertainty is suppressed to a negligible level. Furthermore, we note that the ab initio calculation is performed assuming trial values for m_e/m_p and m_e/m_d , and naturally yields the rotational frequencies (independent of Rydberg constant value), $f_i^{(\text{theor},n)} \approx 1.998... \times 10^{-4}$ atomic units. From these, we compute the scaled, dimensionless values $F_i^{(\text{theor})} = (\mu_{\text{pd}}/m_e) f_i^{(\text{theor},n)}/1$ atomic unit. These have an important dependence on r_p and r_d . The dependence on other fundamental constants is weak, compared with their uncertainties, the largest of which is $\partial \ln F_i^{(\text{theor})} / \partial \ln(m_e/\mu_{\text{pd}}) \approx 4 \times 10^{-3}$. Because of this smallness, it is consistent to use the CODATA 2018 values of the fundamental constants in the computation of $F_i^{(\text{theor})}$. This results in

$$R_\infty m_e (m_p^{-1} + m_d^{-1}) = \frac{f_{\text{spin-avg}}^{(\text{exp})}}{2c f_{\text{spin-avg}}^{(\text{theor})}} = 8,966.20515050(12)_{\text{exp}}(12)_{\text{theor}}(4)_{\text{CODATA2018}} \text{ m}^{-1} \quad (3)$$

($u_r = 2.0 \times 10^{-11}$), where the third uncertainty is due to the proton and deuteron radius uncertainties. The value is in agreement with the CODATA 2018 value of $8,966.20515041(41) \text{ m}^{-1}$ ($u_r = 4.6 \times 10^{-11}$) (Fig. 4). It results from atomic hydrogen spectroscopy (providing R_∞), hydrogen-like ion spin resonance spectroscopy (m_e) and Penning trap mass spectrometry (m_p, m_d). Our result's total uncertainty is smaller by a factor of 2.4 compared with the CODATA 2018 value and ranks among the most precise measurements of a fundamental constant combination.

Owing to the comparatively small CODATA 2018 uncertainty of R_∞ , our improved uncertainty impacts mostly the mass ratio sum $m_e(m_p^{-1} + m_d^{-1})$. Combining equation (3) with the CODATA 2018 values of $R_\infty, m_e/u$ and m_d/u yields the proton mass

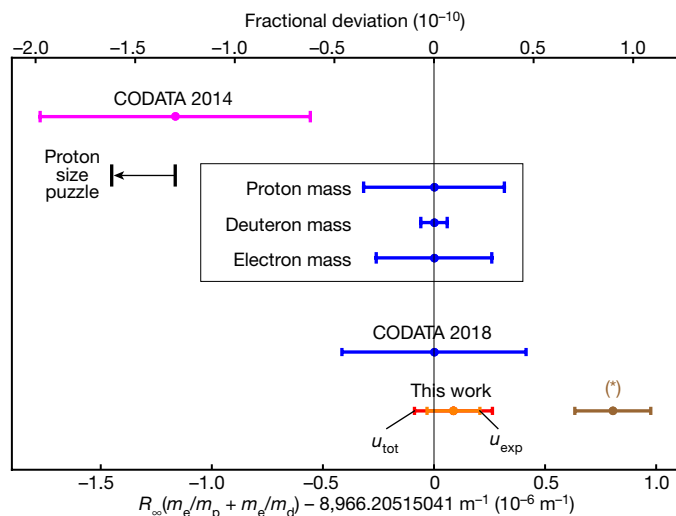


Fig. 4 | Comparison of results of this work with literature values. In the inner box, we plot the error bars for the CODATA 2018 $R_\infty(m_e/m_p + m_e/m_d)$ for the hypothetical cases that the uncertainties of all contributing constants were zero, except for the named constant. The black arrow indicates the shift of the CODATA 2014 value for a change $\Delta R_\infty = -0.00035 \text{ m}^{-1}$ corresponding to the 'proton size puzzle'⁴⁷. The brown data point (*) shows the result of the present work when the CODATA 2014 values of r_p and r_d are used in $f_{\text{spin-avg}}^{(\text{theor})}$ instead of the CODATA 2018 values resulting from muonic hydrogen spectroscopy.

$$m_p/u = 1.007276466605(20)_{\text{exp}}(21)_{\text{theor}}(45)_{\text{CODATA2018}}$$

in excellent agreement with the recent most precise direct measurement⁴⁰

$$m_p/u = 1.007276466598(16)_{\text{stat}}(29)_{\text{syst}}$$

Taking into account a recent Penning trap measurement of m_d/m_p (ref. ⁴¹), we also obtain the proton-to-electron mass ratio

$$m_p/m_e = 1,836.152673449(24)_{\text{exp}}(25)_{\text{theor}}(13)_{\text{CODATA2018, Fink-Myers}}$$

($u_r = 2.0 \times 10^{-11}$) in agreement but approximately two times more accurate than the most precise value, obtained by combining two published measurements in Penning traps^{40,42}: $m_p/m_e = 1,836.152673374(78)_{\text{exp}}$.

Conclusion

The performance of the recently introduced TICTES technique for rotational spectroscopy has been improved by more than two orders in both resolution and accuracy, reaching a fractional FWHM linewidth of 3×10^{-12} and a fractional uncertainty of 1.3×10^{-11} . This vastly higher performance compared with traditional techniques can be of general relevance to the field of precision molecular physics.

Precise measurements of several rotational hyperfine components of HD^+ and suppression of the impact of the limited accuracy of the ab initio theory of the spin structure allowed us to establish agreement between experiment and theory at the 5×10^{-11} level, limited by uncertainties of the CODATA 2018 fundamental constants. To the best of our knowledge, this represents the most accurate test of a molecular physics prediction to date and also provides the most accurate experiment-theory comparison for any three-body quantum system^{2,43-45}. Specifically, we confirmed the combination of the QED contributions of α^5 and α^6 relative order, of the proton finite size contribution and of the deuteron finite size contribution, with uncertainty equal to 0.7% of the

total contribution. A strongly improved upper bound for a new force between a proton and a deuteron was set.

Spin-energy differences were experimentally determined with three orders smaller uncertainty than previously¹². The best (effective) line resolution for spin energy is one order higher and the accuracy is 30 times higher than the benchmark experiment on the spin structure of H_2^+ , which has stood unchallenged for 50 years. The spin-energy predictions were confirmed within the uncertainties of the theory predictions, the smallest uncertainty being 0.7 kHz. As the experimental uncertainties are much lower, the obtained spin-energy data offer new benchmark values for future improved ab initio theory of the spin structure.

We deduced the combinations $R_\infty m_e(m_p^{-1} + m_d^{-1})$ and m_p/m_e of fundamental constants with 2.0×10^{-11} fractional uncertainty, 2.4 and 3.0 times smaller, respectively, than the CODATA 2018 uncertainties. The proton mass in atomic mass units was deduced with the same uncertainty as in CODATA 2018. Interestingly, for the first time, fundamental constants have been determined with competitive uncertainty making use of the rotational motion of a physical system.

Our result also provides independent evidence of the correctness of some of the most precise measurements in atomic and particle physics: Rydberg constant determination via hydrogen spectroscopy, electron mass determination via the bound-electron g -factor, and proton mass and deuteron mass determination via cyclotron motion. Our measurement on a three-body quantum system thus provides an independent link between these one- and two-body systems. The substantial changes introduced in the CODATA 2018 adjustments of the fundamental constants are confirmed. In particular, the predicted HD^+ transition frequency is shifted by 0.063 kHz when the CODATA 2014 proton root-mean-square charge radius and Rydberg constant are replaced by the values deduced from the muonic hydrogen experiment (as in CODATA 2018). Our experimental frequency is consistent with the prediction based on these most recent values, within the combined uncertainties from experiment (0.017 kHz), theory (0.018 kHz) and masses (0.061 kHz).

Beyond the present results, our work has important implications for the near future. First, we suppose that in the spectroscopy of vibrational transitions a similar absolute systematic uncertainty can be achieved as in rotational spectroscopy, because the systematic shifts will not increase substantially with transition frequency. Indeed, the shifts depend on the size of the coefficients of appropriate Hamiltonians, and these coefficients do not vary substantially between the levels. If an optical spectroscopic technique with spectral resolution at the 10-Hz level becomes available, total experimental uncertainties at the 10^{-13} to 10^{-14} level could come into reach. Second, our composite frequency approach obviates the need for a more precise spin-structure theory, both for rotational and vibrational transitions. Therefore, more precise QED calculations of the spin-averaged rotational and vibrational frequencies are both sufficient and well worth pursuing. If this challenging programme is successful, the precision of fundamental constants derived from HD^+ spectroscopy will further improve. Specifically, the combination of rotational and vibrational spectroscopy results and ab initio theory will eventually allow the determination of the fundamental constants R_∞ , m_e/μ_{pd} , r_p and r_d independently rather than in combination, with accuracies competitive with or better than CODATA 2018, and testing QED without limitation by the current determination of the fundamental constants.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2261-5>.

1. Karshenboim, S. G. (ed.) *Precision Physics of Simple Atoms and Molecules* (Springer-Verlag, 2008).
2. Pachucki, K., Patkóš, V. & Yerokhin, V. A. Testing fundamental interactions on the helium atom. *Phys. Rev. A* **95**, 062510 (2017).
3. Leach, C. A. & Moss, R. E. Spectroscopy and quantum mechanics of the hydrogen molecular cation: a test of molecular quantum mechanics. *Annu. Rev. Phys. Chem.* **46**, 55–82 (1995).
4. Roth, B. et al. in *Precision Physics of Simple Atoms and Molecules* (ed. Karshenboim, S. G.) 205–232 (Springer-Verlag, 2008).
5. Wing, W. H., Ruff, G. A., Lamb, W. E. & Spezeski, J. J. Observation of the infrared spectrum of the hydrogen molecular ion HD⁺. *Phys. Rev. Lett.* **36**, 1488–1491 (1976).
6. Arcuni, P. W., Fu, Z. W. & Lundeen, S. R. Energy difference between the ($v = 0, R = 1$) and the ($v = 0, R = 3$) states of H₂⁺, measured with interseries microwave spectroscopy of H₂ Rydberg states. *Phys. Rev. A* **42**, 6950–6953 (1990).
7. Carrington, A., McNab, I. R., Montgomerie-Leach, C. A. & Kennedy, R. A. Vibration-rotation spectroscopy of the HD⁺ ion near the dissociation limit. *Mol. Phys.* **72**, 735–762 (1991).
8. Fu, Z. W., Hessels, E. A. & Lundeen, S. R. Determination of the hyperfine structure of H₂⁺ ($v = 0, R = 1$) by microwave spectroscopy of high- L , $n = 27$ Rydberg states of H₂. *Phys. Rev. A* **46**, R5313–R5316 (1992).
9. Critchley, A. D. J., Hughes, A. N. & McNab, I. R. Direct measurement of a pure rotation transition in H₂⁺. *Phys. Rev. Lett.* **86**, 1725–1728 (2001).
10. Osterwalder, A., Wüest, A., Merkt, F. & Jungen, C. High-resolution millimeter wave spectroscopy and multichannel quantum defect theory of the hyperfine structure in high Rydberg states of molecular hydrogen H₂⁺. *J. Chem. Phys.* **121**, 11810–11838 (2004).
11. Koelmeij, J. C. J., Roth, B., Wicht, A., Ernsting, I. & Schiller, S. Vibrational spectroscopy of HD⁺ with 2-ppb accuracy. *Phys. Rev. Lett.* **98**, 173002 (2007).
12. Bressel, U. et al. Manipulation of individual hyperfine states in cold trapped molecular ions and application to HD⁺ frequency metrology. *Phys. Rev. Lett.* **108**, 183003 (2012).
13. Haase, C., Beyer, M., Jungen, C. & Merkt, F. The fundamental rotational interval of para-H₂⁺ by MQDT-assisted Rydberg spectroscopy of H₂. *J. Chem. Phys.* **142**, 064310 (2015).
14. Biesheuvel, J. et al. Probing QED and fundamental constants through laser spectroscopy of vibrational transitions in HD⁺. *Nat. Commun.* **7**, 10385 (2016).
15. Korobov, V. I., Hilico, L. & Karr, J.-P. Fundamental transitions and ionization energies of the hydrogen molecular ions with few ppt uncertainty. *Phys. Rev. Lett.* **118**, 233001 (2017).
16. Alighanbari, S., Hansen, M. G., Korobov, V. I. & Schiller, S. Rotational spectroscopy of cold and trapped molecular ions in the Lamb–Dicke regime. *Nat. Phys.* **14**, 555–559 (2018).
17. Jefferts, K. B. Hyperfine structure in the molecular ion H₂⁺. *Phys. Rev. Lett.* **23**, 1476–1478 (1969).
18. Schiller, S. & Korobov, V. I. Test of time-dependence of the electron and nuclear masses with ultracold molecules. *Phys. Rev. A* **71**, 032505 (2005).
19. Bakalov, D. & Schiller, S. The electric quadrupole moment of molecular hydrogen ions and their potential for a molecular ion clock. *Appl. Phys. B* **114**, 213–230 (2014); erratum **116**, 777–778 (2014).
20. Karr, J.-P. H₂⁺ and HD⁺: candidates for a molecular clock. *J. Mol. Spectrosc.* **300**, 37–43 (2014).
21. Schiller, S., Bakalov, D. & Korobov, V. I. Simplest molecules as candidates for precise optical clocks. *Phys. Rev. Lett.* **113**, 023004 (2014).
22. Beyer, A. et al. The Rydberg constant and proton size from atomic hydrogen. *Science* **358**, 79–85 (2017).
23. Fleurbaey, H. et al. New measurement of the 1S–3S transition frequency of hydrogen: contribution to the proton charge radius puzzle. *Phys. Rev. Lett.* **120**, 183001 (2018).
24. Bezginov, N. et al. A measurement of the atomic hydrogen Lamb shift and the proton charge radius. *Science* **365**, 1007–1012 (2019).
25. Antognini, A. et al. Proton structure from the measurement of 2S–2P transition frequencies of muonic hydrogen. *Science* **339**, 417–420 (2013).
26. Grémaud, B., Delande, D. & Billy, N. Highly accurate calculation of the energy levels of the H₂⁺ molecular ion. *J. Phys. B* **31**, 383 (1998).
27. Moss, R. E. Energies of low-lying vibration-rotation levels of H₂⁺ and its isotopomers. *J. Phys. B* **32**, L89–L91 (1999).
28. Taylor, J. M., Yan, Z.-C., Dalgarno, A. & Babb, J. F. Variational calculations on the hydrogen molecular ion. *Mol. Phys.* **97**, 25–33 (1999).
29. Tiesinga, E., Mohr, P. J., Newell, D. B. & Taylor, B. N. Values of fundamental physical constants. *NIST* <https://physics.nist.gov/cuu/Constants/index.html> (2019).
30. Wolf, F. et al. Non-destructive state detection for quantum logic spectroscopy of molecular ions. *Nature* **530**, 457–460 (2016).
31. Chou, C. et al. Preparation and coherent manipulation of pure quantum states of a single molecular ion. *Nature* **545**, 203–207 (2017).
32. Schneider, T., Roth, B., Duncker, H., Ernsting, I. & Schiller, S. All-optical preparation of molecular ions in the rovibrational ground state. *Nat. Phys.* **6**, 275–278 (2010).
33. Roth, B., Blythe, P., Wenz, H., Daerr, H. & Schiller, S. Ion-neutral chemical reactions between ultracold localized ions and neutral molecules with single-particle resolution. *Phys. Rev. A* **73**, 042712 (2006).
34. Schiller, S., Roth, B., Lewen, F., Ricken, O. & Wiedner, M. Ultra-narrow-linewidth continuous-wave THz sources based on multiplier chains. *Appl. Phys. B* **95**, 55–61 (2009).
35. Bakalov, D., Korobov, V. I. & Schiller, S. High-precision calculation of the hyperfine structure of the HD⁺ ion. *Phys. Rev. Lett.* **97**, 243001 (2006).
36. Schiller, S. & Korobov, V. I. Canceling spin-dependent contributions and systematic shifts in precision spectroscopy of molecular hydrogen ions. *Phys. Rev. A* **98**, 022511 (2018).
37. Bakalov, D., Korobov, V. I. & Schiller, S. Magnetic field effects in the transitions of the HD⁺ molecular ion and precision spectroscopy. *J. Phys. B* **44**, 025003 (2011); corrigendum **45**, 049501 (2012).
38. Korobov, V. I., Koelmeij, J. C. J., Hilico, L. & Karr, J.-P. Theoretical hyperfine structure of the molecular hydrogen ion at the 1 ppm level. *Phys. Rev. Lett.* **116**, 053003 (2016).
39. Menasian, S. C. & Dehmelt, H. G. High-resolution study of (1/2, 1/2)–(1/2, 3/2) HFS transition in H₂⁺. *Bull. Am. Phys. Soc.* **18**, 408 (1973).
40. Heiße, F. et al. High-precision mass spectrometer for light ions. *Phys. Rev. A* **100**, 022518 (2019).
41. Fink, D. J. & Myers, E. G. Deuteron-to-proton mass ratio from the cyclotron frequency ratio of H₂⁺ to D⁺ with H₂⁺ in a resolved vibrational state. *Phys. Rev. Lett.* **124**, 013001 (2020).
42. Sturm, S. et al. High-precision measurement of the atomic mass of the electron. *Nature* **506**, 467–470 (2014).
43. Pastor, P. C. et al. Absolute frequency measurements of the 2³S₁ → 2³P_{0,1,2} atomic helium transitions around 1083 nm. *Phys. Rev. Lett.* **92**, 023001 (2004).
44. Hori, M. et al. Buffer-gas cooling of antiprotonic helium to 1.5 to 1.7 K, and antiproton-to-electron mass ratio. *Science* **354**, 610–614 (2016).
45. Rengelink, R. J. et al. Precision spectroscopy of helium in a magic wavelength optical dipole trap. *Nat. Phys.* **14**, 1132–1137 (2018).
46. Hori, M. et al. Two-photon laser spectroscopy of antiprotonic helium and the antiproton-to-electron mass ratio. *Nature* **475**, 484–488 (2011).
47. Udem, T. Quantum electrodynamics and the proton size. *Nat. Phys.* **14**, 632–632 (2018); correction **14**, 767 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Experimental procedure

We simultaneously trapped Be^+ and HD^+ ions in a linear RF trap driven at 14.16 MHz (Extended Data Fig. 1). The distance between the trap centre and the RF electrodes was 4.3 mm. For translational cooling of the molecular ions, we laser-cooled the atomic ions with a laser at 313 nm and the HD^+ ions were sympathetically cooled via electrostatic interactions with the Be^+ ions. We estimated the ion secular temperature as about 30 mK. Typically, roughly 10^2 HD^+ ions were trapped together with about 2×10^3 Be^+ ions. The number of trapped HD^+ ions affects the spectral resolution of the rotational transitions, since the Lamb–Dicke regime can only be reached when the ions' displacements in the transverse direction are much smaller than the transition wavelength.

Black-body radiation populates the excited rotational levels of the ground vibrational state until a thermal equilibrium population is reached. We counteracted this by pumping the HD^+ population into the ground rovibrational state using two lasers. They drive the $(0, 2) \rightarrow (1, 1)$ and $(0, 1) \rightarrow (2, 0)$ transitions, and the spontaneous decay from the respective excited states eventually transfers a large fraction of the HD^+ ions in the rovibrational ground state. A quantum cascade laser at 5.48 μm excited the former transition, and a distributed feedback laser at 2.7 μm excited the latter transition.

After rotational cooling, the terahertz radiation was turned on to drive a transition between specific Zeeman components of a specific hyperfine rotational transition. The terahertz wave intensity was controlled with a half-wave plate, a linear polarizer and via the synthesizer output level. A 1.4- μm laser selectively excited molecules from the $(0, 1)$ level to the $(4, 0)$ level. Molecules in this level were rapidly dissociated by a 266-nm laser.

The spectroscopy scheme relies on the ability to determine the relative decrease of the number of trapped HD^+ ions. Resonant excitation of the HD^+ ions' radial secular motion with an auxiliary a.c. electric field couples to the Be^+ ion ensemble, heating it and causing a change in atomic fluorescence. This fluorescence change is approximately proportional to the number of trapped HD^+ ions. Applying the secular excitation before and after the RMPD and calculating the ratio of average fluorescence levels provides the fractional decrease of the number of HD^+ ions. See Extended Data Fig. 2.

As the RMPD process removes HD^+ ions from the trap, repeated loadings are necessary. With one loading of Be^+ , approximately 40 loadings of HD^+ were performed. For each HD^+ loading, typically five spectroscopy cycles were performed. Each cycle lasted 60 s and provided one data point.

The magnetic field was $B_0 \approx 45 \mu\text{T}$, directed along the trap axis, except during rotational spectroscopy/RMPD, when the field was changed to $B \approx 30 \mu\text{T}$ or lower, oriented perpendicular to the trap axis and parallel to the terahertz radiation wave vector (Extended Data Fig. 1). The magnitude and direction of the magnetic field were controlled by three pairs of magnetic coils outside the vacuum chamber.

Owing to the complicated statistics of the ion detection process, we assigned one-half of the FWHM of a line as the statistical uncertainty of a measured transition frequency.

Systematic effects

As a guide to and comparison with the experimental work, the *ab initio* values for various systematic effects were taken from our previous calculations. Explicit values for the Zeeman effect are given in ref. ³⁷, and for the Stark effect in ref. ⁴⁸. The *ab initio* a.c. polarizabilities at the frequency corresponding to the wavelength 266 nm were computed in ref. ¹⁶.

Trap shift. Several systematic shifts are expected to give rise to a quadratic dependence on RF amplitude. These include the micromotion-induced Stark shift⁴⁹, phase-offset-induced Stark shift⁴⁹,

and a.c. Zeeman shift due to an alternating magnetic field at the trap frequency correlated with the electric trap drive.

We therefore measured the dependence of the six lines (including three Zeeman components for line 16 and two Zeeman components for line 19) on the trap RF amplitude. The typical values chosen for the RF amplitude were 150 V, 180 V and 245 V. The precise RF amplitude value for each measurement was determined by measuring the radial secular frequency of Be^+ . See Extended Data Fig. 3 for an example of the frequency shift when varying the trap's RF field amplitude. Fits, assuming quadratic dependence, furnish the correction to be applied for obtaining each line's extrapolated frequency for zero RF amplitude. The theory of the Stark shift⁴⁸ predicts shifts of the same sign (positive) and of similar value for all components considered here. The experimental data are consistent with this prediction.

Zeeman shift. Both the linear and quadratic Zeeman shift coefficients vary substantially among Zeeman components and hyperfine components (compare, for example, lines 16 and 19 in Fig. 2). The frequency splitting of the two Zeeman components 16_{\pm} together with the theoretical linear Zeeman splitting coefficient ($7.98 \text{ kHz } \mu\text{T}^{-1}$ (ref. ³⁷)) allows the determination of the (time- and ensemble-averaged) magnetic field affecting the molecular ions. For the data shown in Fig. 2, the nominal magnetic field $B_{\text{nom}} = 2.98(3) \times 10^{-5} \text{ T}$ is consistent with the value deduced using spectroscopy of the co-trapped beryllium ions⁵⁰. The observed linewidth of the 16_{\pm} Zeeman components indicates that the magnetic field is homogeneous to at least 1 part in 30 over the molecule sample.

We measured the frequencies at three different values of magnetic field, for RF amplitudes close to the nominal value of 190 V. Since the RF amplitude varied slightly for the individual measurements, each measured frequency was corrected for the trap shift.

To obtain the $B \rightarrow 0$ extrapolated frequency, $f_i^{(\text{exp})}$, for each line, we fitted to the measured line frequencies $f_i^{(\text{exp})}(B)$ the sum of $f_i^{(\text{exp})}$ plus a quadratic-in- B and/or linear-in- B dependence, depending on the type of Zeeman component. As an accurate measure of the magnetic field, we used the splitting $f_{16_{-}} - f_{16_{+}}$. For $m_F = 0 \rightarrow m'_F = 0$ Zeeman components, we assumed a quadratic-in- B dependence. For the two components 19_{\pm} and for the two components 16_{\pm} , we allowed for independent linear-in- B shift coefficients $\alpha_{i,+}$, $\alpha_{i,-}$. For $f_{16_{-}}, f_{16_{+}}$, we added to the fit functions the quadratic Zeeman shift predicted by theory. From the fits, we found that the 'positive' and 'negative' shift coefficients of a given line are close: $\alpha_{19,-} \approx \alpha_{19,+}$ and $\alpha_{16,-} \approx \alpha_{16,+}$.

The input data for the magnetic-field dependence fit are the trap-field-extrapolated line frequencies. The reported uncertainty of each $f_i^{(\text{exp})}$ contains both the uncertainty of the magnetic-field extrapolation and the uncertainty due to the trap-field extrapolation.

The magnetic field is produced by three solenoids. They were characterized with a magnetic probe before closing the vacuum chamber. We find the field value deduced from the solenoids' currents agrees with the value deduced from the splitting $f_{16_{-}} - f_{16_{+}}$, within the experimental uncertainty of the former.

Trap-induced a.c. Zeeman shift. This effect would show up as a variation of the splitting between two Zeeman components with the trap RF amplitude. The 19_{\pm} components were measured at 245 V and 154 V, at the nominal magnetic field. Their frequency difference did not change, indicating a negligible a.c. Zeeman shift.

Light shift due to cooling laser. The 313-nm cooling laser permanently irradiates the ion cluster, including when the terahertz wave is on. Its nominal power is 100 μW and the beam radius is 0.25 mm. We measured the effect of a change of the 313-nm laser intensity on $f_{16_{-}}$. No shift was discernible at the 10-Hz level upon increase of the power by a factor of four.

We computed the scalar, tensor and vector polarizabilities of the rovibrational levels at $\lambda = 313 \text{ nm}$ using high-precision variational

wavefunctions, similar to ref. ⁴⁸, obtaining $\alpha_s(\nu=0, L=1) = 3.5054$, $\alpha_t(\nu=0, L=1) = -0.955$, $\alpha_s(\nu=0, L=0) = 3.4961$ and $\alpha_t(\nu=0, L=0) = 0$, in atomic units. The vector polarizabilities are negligible. The computed light shift is of the order of 0.01 Hz. We therefore set the correction due to the 313-nm wave intensity to zero.

Line pulling. We have no observational evidence that Zeeman components, or micromotion-induced sidebands of other hyperfine components, could affect the measured transitions. The small linewidths of the measured transitions are important in this respect. We did not observe any change of $f_{16'}$, f_{16} and f_{16_0} at the 10-Hz level upon a 500-Hz change of the trap frequency.

d.c. offsets. For every measurement reported in the manuscript, the HD^+ ions are located along the symmetry axis of the Be^+ ion cluster. An offset of 10 V was applied to an electrode to displace the beryllium crystal by about 100 μm from the trap axis along the radial direction. We observed that this offset potential does not have an effect on the position of the HD^+ ions, as also found in molecular dynamics simulations¹⁶. We measured the frequency shift of f_{19} caused by this offset potential to be 1(10) Hz. Possible day-to-day variations of the trap compensation voltage are a small fraction of the applied offset. Therefore, the size and uncertainty resulting from these variations are negligible.

Light shift due to the two REMP lasers. The shift due to the 1.4- μm laser and 266-nm laser waves present during spectroscopy has been determined by performing spectroscopy in a different mode, alternating terahertz irradiation and REMP laser irradiation. The shift has been measured for all lines and all Zeeman components discussed here. The shifts are smaller than or equal to 0.039(17) kHz in absolute value. The measured shifts and their uncertainties are used as corrections.

Other shifts. According to theoretical calculations, the black-body radiation shift⁴⁸ and the molecular electric quadrupole shift⁵¹ can be neglected at the present level of accuracy.

Data analysis

Extrapolation of the measured frequencies to zero magnetic field and zero trap amplitude is done by a standard least-squares method. Standard formulae for the propagation of uncertainties are applied.

Spin coefficients, their uncertainties, and sensitivity of the transition frequencies to the spin coefficients

To allow for an accurate comparison between experiment and ab initio theory, we performed a substantially more accurate computation of the spin-structure coefficients of HD^+ compared with our earlier work³⁵. We extended the approach developed in ref. ³⁸ and the relevant matrix elements were calculated to ten significant digits. Values of the two spin-structure coefficients for the lower level, \mathcal{E}_4 and \mathcal{E}_5 , and the nine coefficients for the upper level, $\mathcal{E}'_1, \dots, \mathcal{E}'_9$ are reported in the Extended Data Table 1. Using these coefficients in the diagonalization of the spin-structure Hamiltonian of ref. ³⁵, we obtain the spin frequencies $f_{\text{spin},i}$ (Extended Data Table 1).

The largest spin-structure coefficients, $\mathcal{E}_4, \mathcal{E}'_4, \mathcal{E}_5$ and \mathcal{E}'_5 , have theoretical fractional uncertainties of approximately $\mathcal{E}_4 \approx \mathcal{E}_5 \approx 1 \times 10^{-6} = \varepsilon_F$. This estimate is confirmed by comparison of the theoretical predictions of the molecular ion H_2^+ , calculated with the same theoretical approach, with the experimental results of refs. ^{17,39}. For a given vibrational level, the rotational dependence of the neglected terms in \mathcal{E}_4 and \mathcal{E}_5 is nearly zero, because these are contact terms determined by the electronic wave function, which depends very weakly on N . This allows us to assume that the neglected terms in $(\mathcal{E}_4, \mathcal{E}'_4)$ and in $(\mathcal{E}_5, \mathcal{E}'_5)$ are essentially equal, respectively.

Under this assumption, the theory uncertainty of a spin frequency due to these coefficients $k = 4, 5$ is set to $u_k = |\gamma'_{i,k} \mathcal{E}'_k - \gamma_{i,k} \mathcal{E}_k| \varepsilon_F$, where

$\gamma_{i,k} = -\partial f_{\text{spin},i} / \partial \mathcal{E}_k$ is the derivative of the spin energy of the lower quantum state involved in the transition i with respect to the spin coefficient \mathcal{E}_k , and $\gamma'_{i,k} = \partial f_{\text{spin},i} / \partial \mathcal{E}'_k$ is defined analogously for the upper state. The values of the derivatives are presented in Extended Data Table 1.

The spin Hamiltonian coefficients $\mathcal{E}_4 \approx \mathcal{E}'_4$ and $\mathcal{E}_5 \approx \mathcal{E}'_5$ are similar for the two rotational states, and because the transitions studied here are those between similar spin states, for which $G_1 = G'_1, G_2 = G'_2$, the spin frequencies are small, $|f_{\text{spin},i}| \ll \mathcal{E}_4, \mathcal{E}_5, \mathcal{E}'_4, \mathcal{E}'_5$ and the sensitivities are similar, $\gamma'_{i,k} \approx \gamma_{i,k}$. Therefore, we benefit from important reduction of the theory uncertainties u_4 and u_5 contributed by these four coefficients. Even in the least favourable case, line 14, the uncertainty contribution is less than or equal to $u_4 + u_5 \approx 14 \text{ Hz}$ (1×10^{-11}), that is, negligible compared with the following contributions.

A second set of coefficients, $\mathcal{E}'_2, \mathcal{E}'_6$ and \mathcal{E}'_7 , are one to three orders smaller in magnitude, and have estimated fractional uncertainties of $\varepsilon_1 \approx \varepsilon_6 \approx \varepsilon_7 \approx \alpha^2 = \varepsilon_0 \approx 5 \times 10^{-5}$. Their absolute uncertainties, 1.5 kHz to 0.06 kHz, are at a relevant level. They enter the spin-structure frequency uncertainty with contributions $u_k = |\mathcal{E}'_k| \varepsilon_0$.

The fractional uncertainties of the coefficients $\mathcal{E}'_2, \mathcal{E}'_3, \mathcal{E}'_8$ and \mathcal{E}'_9 are similar to ε_0 , but are not relevant at the present experimental accuracy level because the coefficients themselves are much smaller than the others.

As the details of the theory errors are unknown, the total uncertainty of the spin frequencies is set conservatively as the sum over all u_k (instead of the root sum of squares).

The sensitivities γ are obtained by first computing the eigenvalues $E_{\text{spin},i}$ and $E'_{\text{spin},i}$ of the Hamiltonian analytically and then computing analytically their derivatives with respect to the individual coefficients \mathcal{E}_k and \mathcal{E}'_k . These derivatives are then evaluated for the set of current theory values for \mathcal{E}_k and \mathcal{E}'_k .

Fit of the spin Hamiltonian coefficients

From the six measured transitions, we can derive information about the spin Hamiltonian coefficients and about the true spin-averaged frequency. Under the previous assumption of equal theory errors for $(\mathcal{E}_4, \mathcal{E}'_4)$ and for $(\mathcal{E}_5, \mathcal{E}'_5)$, there are six remaining important quantities $(\mathcal{E}'_1, \mathcal{E}'_4, \mathcal{E}'_5, \mathcal{E}'_6, \mathcal{E}'_7 \text{ and } f_{\text{spin-avg}})$, and they can be solved for using a set of equations in which the experimental frequencies are equal to the corresponding theoretical frequencies, allowing for small deviations from the nominal values. We find $\mathcal{E}'_1^{(\text{fit})} - \mathcal{E}'_1^{(\text{theor})} = 0.32(20) \text{ kHz}$ where the uncertainty is smaller than the theory uncertainty, $\varepsilon_F \mathcal{E}'_1 \approx 1.6 \text{ kHz}$.

Furthermore, $\mathcal{E}'_6^{(\text{fit})} - \mathcal{E}'_6^{(\text{theor})} = 0.5(9) \text{ kHz}$, $\mathcal{E}'_7^{(\text{fit})} - \mathcal{E}'_7^{(\text{theor})} = -0.3(4) \text{ kHz}$ and $f_{\text{spin-avg}}^{(\text{fit})} - f_{\text{spin-avg}}^{(\text{theor})} = -0.05(22) \text{ kHz}$. The shown uncertainties result from the experimental errors and the theory error of $f_{\text{spin-avg}}^{(\text{theor})}$; the theory errors of $\mathcal{E}'_2, \mathcal{E}'_3, \mathcal{E}'_8$ and \mathcal{E}'_9 make negligible contributions. The deviations of \mathcal{E}'_4 and \mathcal{E}'_5 from the nominal values cannot be determined precisely (an aspect that is intrinsic to the favoured transitions), but are consistent with zero.

Composite frequencies

The coefficients of the composite frequency given in the main text are:

$$\begin{aligned} b_{12} &= 0.0863720 \dots, & b_{14} &= 0.1456348 \dots, & b_{16} &= 0.2516111 \dots, \\ b_{19} &= 0.2442792 \dots, & b_{20} &= 0.1328074 \dots, & b_{21} &= 0.1392955 \dots \end{aligned}$$

We consider alternative composite frequencies. One alternative ansatz for finding a composite frequency is to impose the 'insensitivity conditions' $0 = \partial f_c^{(\text{theor})} / \partial \mathcal{E}_{k_a} = \sum_i b_i \gamma_{i,k_a}, 0 = \partial f_c^{(\text{theor})} / \partial \mathcal{E}'_{k_\beta}$ for a suitable subset $\{k_a, k_\beta\}$ of spin Hamiltonian coefficients. As discussed above, if we assume correlated errors for the pair $(\mathcal{E}_4, \mathcal{E}'_4)$ and $(\mathcal{E}_5, \mathcal{E}'_5)$, then the largest theory uncertainties arise from $\mathcal{E}'_1, \mathcal{E}'_6$ and \mathcal{E}'_7 . Four experimentally measured transitions are sufficient to satisfy the three insensitivity conditions for these three coefficients. The normalization condition

is easily imposed in addition. Considering, for example, the lines 14, 16, 19 and 21, the resulting uncertainty from hyperfine theory is $u(f_{\text{spin,c}}^{\text{(theor)}}) \approx 2$ Hz, much smaller than the uncertainty of the spin-averaged frequency $u(f_{\text{spin-avg}}^{\text{(theor)}}) \approx 0.02$ kHz. Thus, the composite frequency has a substantially reduced theory uncertainty compared with those of the individual hyperfine transitions. $f_c^{\text{(theor)}}$ is then also numerically close to $f_{\text{spin-avg}}^{\text{(theor)}}$, $f_c^{\text{(theor)}} \approx f_{\text{spin-avg}}^{\text{(theor)}} + 2,232$ kHz. With more available transitions we can impose additional conditions.

A second alternative composite frequency is as follows. As in the main text, we consider a composite frequency that minimizes the spin-coefficients-related uncertainty. If we assume correlated \mathcal{E} errors, the linear combination of only three lines, $f_c = b_{14}f_{14} + b_{16}f_{16} + (1 - b_{14} - b_{16})f_{21}$, yields an uncertainty of 3 Hz (2.4×10^{-12}). As in the first alternative, this uncertainty is also much smaller than $u(f_{\text{spin-avg}}^{\text{(theor)}})$. The coefficients are $b_{14} = 0.0814\dots$, $b_{16} = 0.615\dots$ and $f_c^{\text{(theor)}} = f_{\text{spin-avg}}^{\text{(theor)}} + 1,524.23$ kHz. Such optimal solutions exist independently of the concrete values of the estimated theory uncertainties of the \mathcal{E} coefficients: if the assumed fractional uncertainties ε_i are doubled, a solution is obtained whose theory uncertainty is correspondingly larger, 6 Hz. The relationship between the solution $f_c^{\text{(theor)}}$ and the cancellation conditions is that the determinant of the sensitivity matrix $\Gamma_{i,k} = \gamma'_{i,k} = \partial f_i^{\text{(theor)}} / \partial \varepsilon'_k$ (where $i = \{14, 16, 21\}$ and $k = \{1, 6, 7\}$), is close to zero (about 0.008). This implies that these three transitions are nearly linearly dependent and allow for a composite frequency that nearly satisfies the cancellation conditions (and the normalization condition).

If the correlation assumption is not made, the optimum composite frequency based on lines 14, 16 and 21 yields a comparatively large spin-energy uncertainty of 0.22 kHz. For this reason, in the main text, we determined the composite frequency based on six lines.

A third example is the composite frequency based on the five lines 14, 15, 16, 19 and 20: it yields a theory uncertainty $u(f_{\text{spin,c}}^{\text{(theor)}}) \approx 3$ Hz.

Finally, an example of composite frequency for a vibrational transition is the following. For the transition $(v = 0, N = 0) \rightarrow (v' = 1, N' = 1)$ the six lines 14, 15, 16, 19, 20 and 21 yield a composite frequency with theory uncertainty $u(f_{\text{spin,c}}^{\text{(theor)}}) \approx 2$ Hz. This is only 3×10^{-14} relative to the vibrational transition frequency $f_{\text{spin-avg}} \approx 58.6$ THz.

Fifth force bound

Given the present results, the 95% confidence limit to the strength of the fifth force, $\beta_{\text{max}}(\lambda)$, is approximately given by

$$N_1 N_2 |\Delta Y(\lambda)| \beta_{\text{max}}(\lambda) \approx 2 h u_{\text{tot}}(f_{\text{rot}}),$$

$$u_{\text{tot}}(f_{\text{rot}})^2 = u(f_{\text{spin-avg}}^{\text{(exp)}})^2 + u(f_{\text{spin-avg}}^{\text{(theor)}})^2 + u_{\text{CODATA2018}}(f_{\text{spin-avg}}^{\text{(theor)}})^2$$

Here, $\Delta Y(\lambda)$ is obtained numerically from perturbation theory as the difference of the expectation value of $R^{-1} \exp(-R/\lambda)$ in the two rotational states, where R is the internuclear separation divided by 1 atomic unit, and λ , N_1 and N_2 were defined in Fig. 3.

We have also obtained an analytical approximate expression

$$\beta_{\text{max}}(\lambda) \approx 2 \frac{u_{\text{tot}}(f_{\text{rot}})}{f_{\text{rot}}} \frac{e^{R_e/\lambda}}{2 N_1 N_2 (1 + R_e/\lambda)} \frac{R_e E_{\text{vib}}^2}{E_{\text{rot}}} \quad (4)$$

where R_e is the equilibrium separation, and $E_{\text{rot}} = f_{\text{rot}}/2cR_\infty$ and E_{vib} are the fundamental rotational transition energy and fundamental vibrational transition energy, respectively. They are all normalized

to the respective atomic unit. The previous bounds on β are also discussed in ref. ⁵².

Electric quadrupole moment of the deuteron

We deduce a value for the electric quadrupole moment of the deuteron, Q_d . The tensor interaction between Q_d and the electric field gradient within the HD⁺ molecule³⁵ contributes to the hyperfine structure. It is quantified by the spin Hamiltonian coefficient $\mathcal{E}'_9 = 5.666$ kHz $\propto Q_d$. The ratio \mathcal{E}'_9/Q_d is available from our theory with small fractional uncertainty $\varepsilon_9 \approx 5 \times 10^{-5}$. The frequencies of the rotational transition components are sensitive to \mathcal{E}'_9 to varying degrees, quantified by $\gamma'_{i,9}$ (see Extended Data). We therefore consider a composite frequency $f'_c = \sum_i a_i f_i$ that suppresses the spin-averaged frequency, and thus all QED contributions, by imposing $\sum_i a_i = 0$. We determine the weight set $\{a_i\}$ that maximizes the sensitivity-to-uncertainty ratio $|\partial f'_c / \partial \mathcal{E}'_9|^2 / (u(f_c^{\text{(theor)}})^2 + u(f_c^{\text{(exp)}})^2)$. We find $a_{12} = -0.2165167$, $a_{14} = 0.6508068$, $a_{16} = -0.9098989$, $a_{19} = -0.9738303$ and $a_{20} = -0.1153690$.

From the comparison of $f_c^{\text{(theor)}}$ and $f_c^{\text{(exp)}}$, we then deduce $Q'_d = 0.282(4)$ fm². It is consistent with the reference value $Q_d = 0.28578(3)$ fm², obtained from RF spectroscopy of neutral D₂ and theory⁵³. The precision is expected to improve with progress in MHI spin-structure theory and experimental precision.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

48. Schiller, S., Bakalov, D., Bekbaev, A. K. & Korobov, V. I. Static and dynamic polarizability and the Stark and blackbody-radiation frequency shifts of the molecular hydrogen ions H₂⁺, HD⁺, and D₂⁺. *Phys. Rev. A* **89**, 052521 (2014).
49. Berkeland, D. J., Miller, J. D., Bergquist, J. C., Itano, W. M. & Wineland, D. J. Minimization of ion micromotion in Paul trap. *J. Appl. Phys.* **83**, 5025–5033 (1998).
50. Shen, J., Borodin, A. & Schiller, S. A simple method for characterization of the magnetic field in an ion trap using Be⁺ ions. *Eur. Phys. J. D* **68**, 359 (2014).
51. Bakalov, D. & Schiller, S. The electric quadrupole moment of molecular hydrogen ions and their potential for a molecular ion clock. *Appl. Phys. B* **114**, 213–230 (2014); corrigendum **116**, 777–778 (2014).
52. Salumbides, E. J., Ubachs, W. & Korobov, V. I. Bounds on fifth forces at the sub-Å length scale. *J. Mol. Spectrosc.* **300**, 65–69 (2014).
53. Pavanello, M., Tung, W.-C. & Adamowicz, L. Determination of deuteron quadrupole moment from calculations of the electric field gradient in D₂ and HD. *Phys. Rev. A* **81**, 042526 (2010).

Acknowledgements We thank M. G. Hansen for assistance with optimization of the apparatus and J.-Ph. Karr for checking theoretical expressions. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 786306, 'PREMOL'), and from the Deutsche Forschungsgemeinschaft in project Schi 431/23-1. S.A. acknowledges a fellowship of the Prof.-W.-Behmenburg-Schenkung. V.I.K. acknowledges support from the Russian Science Foundation under grant number 18-12-00128.

Author contributions S.A. and G.S.G. performed the measurements and analysed data, F.L.C. contributed to the measurements. S.A. developed and maintained the apparatus. V.I.K. performed the ab initio calculations, S.S. performed data and theoretical analyses, prepared the manuscript and supervised the work. All authors contributed to discussion and manuscript editing.

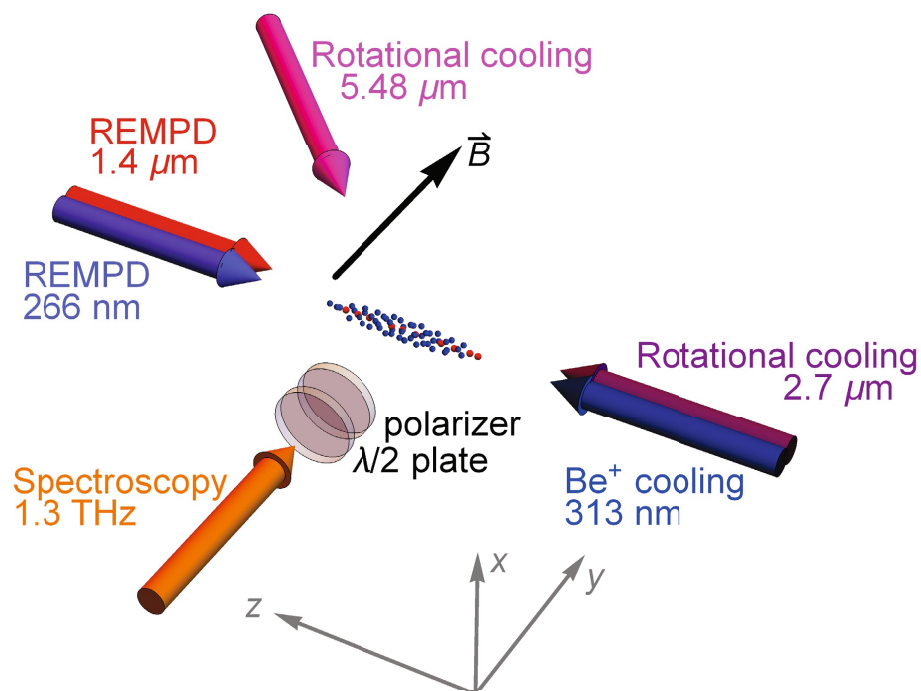
Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.S.

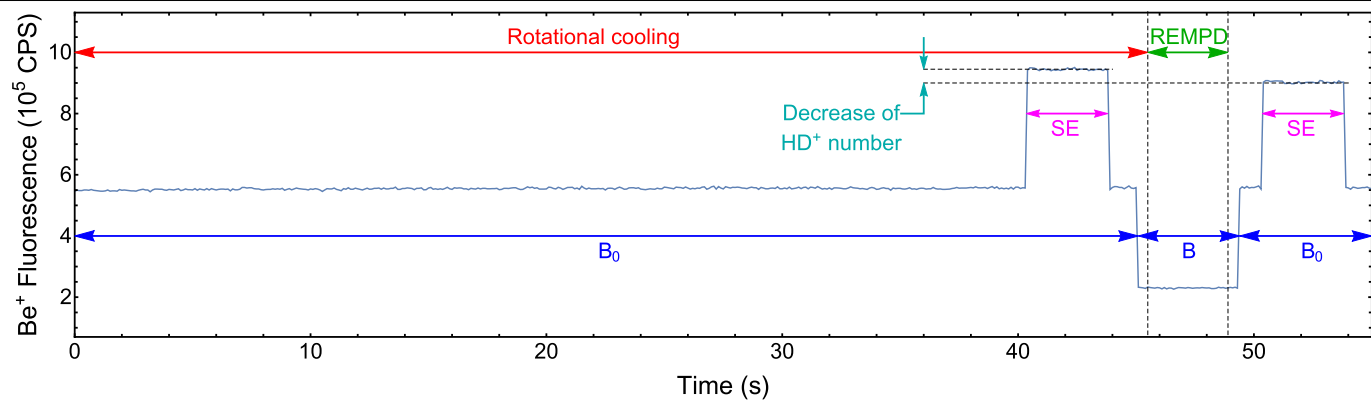
Peer review information Nature thanks Brian Odom, Richard Thompson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



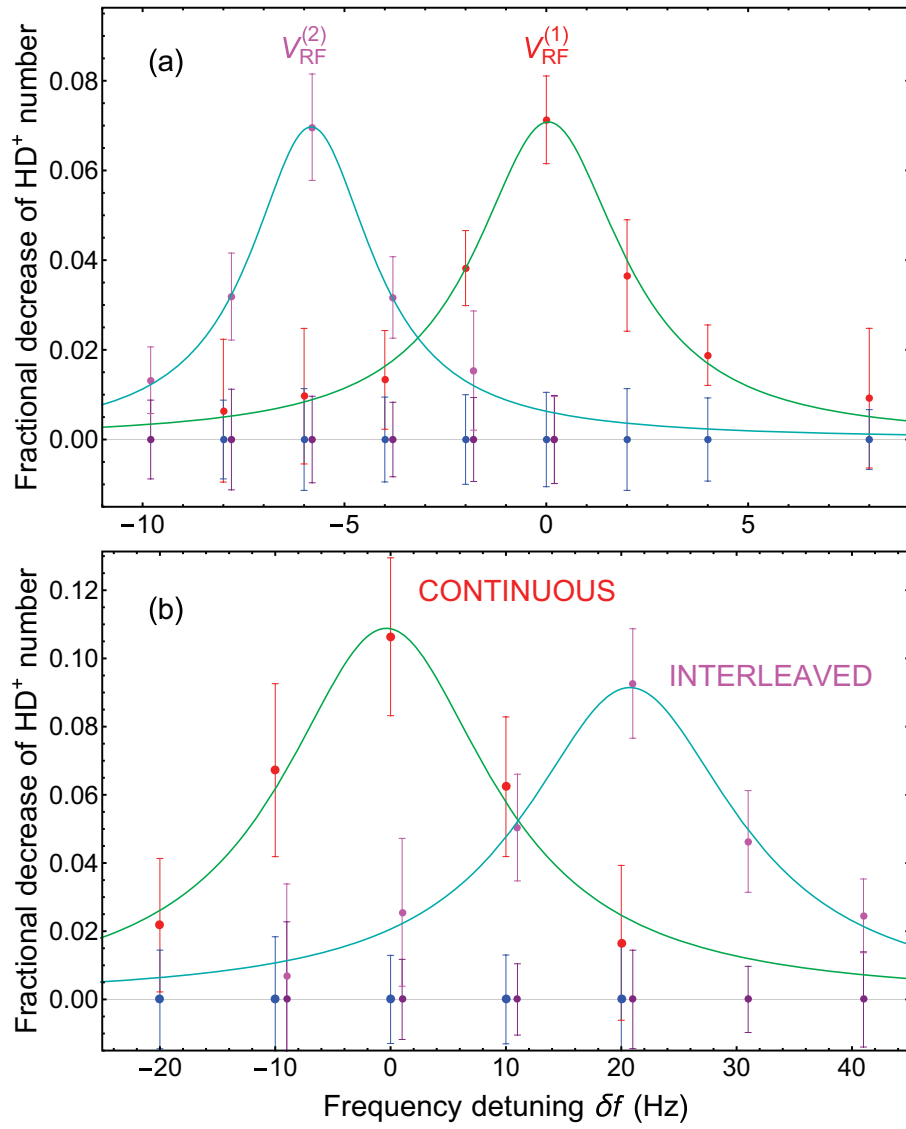
Extended Data Fig. 1 | Conceptual view of the arrangement used for high-resolution spectroscopy of HD^+ using TICTES. The spectroscopy wave (1.3 THz) crosses the ion cluster perpendicular to its long axis, enabling spectroscopy in the Lamb–Dicke regime. The ion cluster comprises atomic Be^+ ions (blue dots) and HD^+ molecular ions (red dots). The indicated laser beams

implement the Doppler cooling of Be^+ ions (313 nm), rotational cooling of HD^+ (2.7 μm and 5.48 μm) and detection by REMPD (266 nm and 1.4 μm). The magnetic field \mathbf{B} lifts the degeneracy of Zeeman sublevels during terahertz spectroscopy. The polarizer and the half-wave plate enable adjustment of the polarization and intensity of the terahertz radiation.



Extended Data Fig. 2 | Beryllium ion fluorescence during one preparation-spectroscopy cycle. Spectroscopy (terahertz wave on) occurs during the interval marked 'REMPD'. Beryllium laser cooling is on all the time. SE, secular

excitation. B, a magnetic flux strength B is applied during REMPD. B_0 , a strength B_0 is applied for rotational laser cooling. CPS, counts per second. The signal obtained from the spectroscopy cycle is indicated in cyan.



Extended Data Fig. 3 | Systematic shifts of the Zeeman component 19, of the rotational hyperfine transition line 19. a, The trap's amplitude is decreased by 2.5 V from $V_{\text{RF}}^{(1)}$ to $V_{\text{RF}}^{(2)}$. The FWHM linewidth is 4 Hz, corresponding to 3×10^{-12} fractional FWHM. **b,** The light shift induced by the 266 nm and 1.4 μm

dissociation lasers, determined by comparing two spectroscopy modes. 'Continuous' indicates that the lasers are on when the terahertz radiation is applied. 'Interleaved' indicates that the lasers and terahertz radiation are on alternatingly.

Extended Data Table 1 | Spin Hamiltonian coefficients, spin-structure frequencies and spin-frequency derivatives

		\mathcal{E}_1'	\mathcal{E}_2'	\mathcal{E}_3'	\mathcal{E}_4'	\mathcal{E}_5'	\mathcal{E}_6'	\mathcal{E}_7'	\mathcal{E}_8'	\mathcal{E}_9'	\mathcal{E}_4	\mathcal{E}_5
		31.98465	−0.03134	−0.004810	924.56943	142.16092	8.61111	1.32177	−0.003057	0.005666	925.39588	142.28781
Line i	$f_{\text{spin},i}^{(\text{theor})}$	$\gamma'_{i,1}$	$\gamma'_{i,2}$	$\gamma'_{i,3}$	$\gamma'_{i,4}$	$\gamma'_{i,5}$	$\gamma'_{i,6}$	$\gamma'_{i,7}$	$\gamma'_{i,8}$	$\gamma'_{i,9}$	$\gamma_{i,4}$	$\gamma_{i,5}$
12	−33.20866	−0.569	−0.559	−1.718	0.250	0.425	0.039	−3.347	−3.284	−2.944	0.250	0.500
14	−9.07415	−0.429	−0.386	0.654	0.249	−0.908	−1.028	0.834	0.721	−0.502	0.250	−1.000
16	−2.13496	−0.107	0.111	0.995	−0.734	−0.184	0.010	0.143	−0.157	−0.493	−0.737	−0.169
19	10.07468	0.500	0.500	1.000	0.250	0.500	−0.500	−1.000	−1.000	−0.500	0.250	0.500
20	11.73591	−0.225	−0.265	−0.510	0.250	0.500	1.734	3.439	3.559	1.764	0.250	0.500
21	11.78771	0.332	0.154	0.514	0.234	−0.315	0.255	−0.582	−0.402	0.229	0.237	−0.331

\mathcal{E}'_k (\mathcal{E}_k) are the updated coefficients of the spin Hamiltonian³⁵ of the upper (lower) rotational level, in MHz. $f_{\text{spin},i}^{(\text{theor})}$ are theoretical spin frequencies in MHz. γ are the dimensionless sensitivities of the spin frequencies to the spin Hamiltonian coefficients. $\gamma'_{i,k} = \partial f_{\text{spin},i}^{(\text{theor})} / \partial \mathcal{E}'_k$ refer to the upper state and $\gamma_{i,k} = -\partial f_{\text{spin},i}^{(\text{theor})} / \partial \mathcal{E}_k$ to the lower state. The entries for line 19 are decimal representations of rational values (see equation (6) in ref. ³⁷). Note that because of the tracelessness of the spin Hamiltonian³⁶, $\sum_i d_i \gamma_{i,k} = 0$ and $\sum_i d'_i \gamma'_{i,k} = 0$, where $d_i = (2F(i) + 1)/36$ and $d'_i = (2F'(i) + 1)/36$ are the degeneracies of the respective spin states, and the sum is over the ten favoured transitions $i = 12, \dots, 21$.

Spin squeezing of 10^{11} atoms by prediction and retrodiction measurements

<https://doi.org/10.1038/s41586-020-2243-7>

Received: 11 July 2019

Accepted: 26 February 2020

Published online: 13 May 2020

 Check for updates

Han Bao¹, Junlei Duan¹, Shenchao Jin¹, Xingda Lu¹, Pengxiong Li¹, Weizhi Qu¹, Mingfeng Wang^{1,2}, Irina Novikova³, Eugeny E. Mikhailov³, Kai-Feng Zhao⁴, Klaus Mølmer⁵✉, Heng Shen^{6,7}✉ & Yanhong Xiao^{1,6}✉

The measurement sensitivity of quantum probes using N uncorrelated particles is restricted by the standard quantum limit¹, which is proportional to $1/\sqrt{N}$. This limit, however, can be overcome by exploiting quantum entangled states, such as spin-squeezed states². Here we report the measurement-based generation of a quantum state that exceeds the standard quantum limit for probing the collective spin of 10^{11} rubidium atoms contained in a macroscopic vapour cell. The state is prepared and verified by sequences of stroboscopic quantum non-demolition (QND) measurements. We then apply the theory of past quantum states^{3,4} to obtain spin state information from the outcomes of both earlier and later QND measurements. Rather than establishing a physically squeezed state in the laboratory, the past quantum state represents the combined system information from these prediction and retrodiction measurements. This information is equivalent to a noise reduction of 5.6 decibels and a metrologically relevant squeezing of 4.5 decibels relative to the coherent spin state. The past quantum state yields tighter constraints on the spin component than those obtained by conventional QND measurements. Our measurement uses 1,000 times more atoms than previous squeezing experiments^{5–10}, with a corresponding angular variance of the squeezed collective spin of 4.6×10^{-13} radians squared. Although this work is rooted in the foundational theory of quantum measurements, it may find practical use in quantum metrology and quantum parameter estimation, as we demonstrate by applying our protocol to quantum enhanced atomic magnetometry.

Measurements constitute the foundations of physical science. The aim of high-precision metrology is to reduce uncertainties and draw as accurate conclusions as possible from measurement data¹. Quantum systems are described by wave functions or density matrices, which yield probabilistic measurement outcomes. For a continuously monitored system, the well established theory of quantum trajectories employs stochastic master equations to describe the evolution with time of the density matrix $\rho(t)$, which is governed by the system Hamiltonian, dissipation, and effects associated with the measurements². For Gaussian states and operations, the theory is simplified to equations for mean values and covariances, equivalent to classical Kalman filter theory¹¹.

By knowing the value of $\rho(t)$, we can predict the outcome of a subsequent measurement on the system, and if QND probing has led to a state with reduced uncertainty on a specific observable, we may thus make an improved prediction of the subsequent measurement. Also, later measurements will have outcomes correlated with the present and previous ones; in the same way that daily life experience teaches us about past events and facts, one may ask if it is possible in a quantum experiment to obtain more knowledge about a quantum state by using both earlier and later observations on a system. Such retrodiction was

initially introduced in the context of pre- and post-selection under projective measurements¹² and in the theory of weak value measurements¹³, whereas the idea of a complete description of a quantum system at any time during a sequence of measurements¹⁴ has found a general dynamical formulation in the so-called past quantum state (PQS)^{3,4}. The PQS provides the probability distribution of the outcome of any general measurement on a quantum system at time t , conditioned on our knowledge about the system that is obtained by measurements performed both before and after t . The PQS has been demonstrated to yield better predictions than the usual conditional density matrix in trajectory simulations of the photon number evolution in a cavity¹⁵, the excitation and emission dynamics of a superconducting qubit¹⁶ and the motional state of a mechanical oscillator¹⁷.

Here we show that the PQS elements of the quantum trajectory description could further improve already precise measurements with vapour cells for magnetometry^{18–20}, fundamental symmetry tests^{21,22} and gravitational-wave detection²³. In particular, we show that for a metrologically relevant macroscopic atomic spin system, the standard quantum limit determined by the atom projection noise can be surpassed by conditioning the measurement result on previous and

¹Department of Physics, State Key Laboratory of Surface Physics and Key Laboratory of Micro and Nano Photonic Structures, Ministry of Education, Fudan University, Shanghai, China.

²Department of Physics, Wenzhou University, Zhejiang, China. ³Department of Physics, College of William and Mary, Williamsburg, VA, USA. ⁴Applied Ion Beam Physics Laboratory, Key Laboratory of the Ministry of Education, and Institute of Modern Physics, Fudan University, Shanghai, China. ⁵Department of Physics and Astronomy, Aarhus University, Aarhus, Denmark. ⁶State Key Laboratory of Quantum Optics and Quantum Optics Devices, Shanxi University, Taiyuan, China. ⁷Clarendon Laboratory, University of Oxford, Oxford, UK. ✉e-mail: moelmer@phys.au.dk; heng.shen@physics.ox.ac.uk; yxiao@fudan.edu.cn

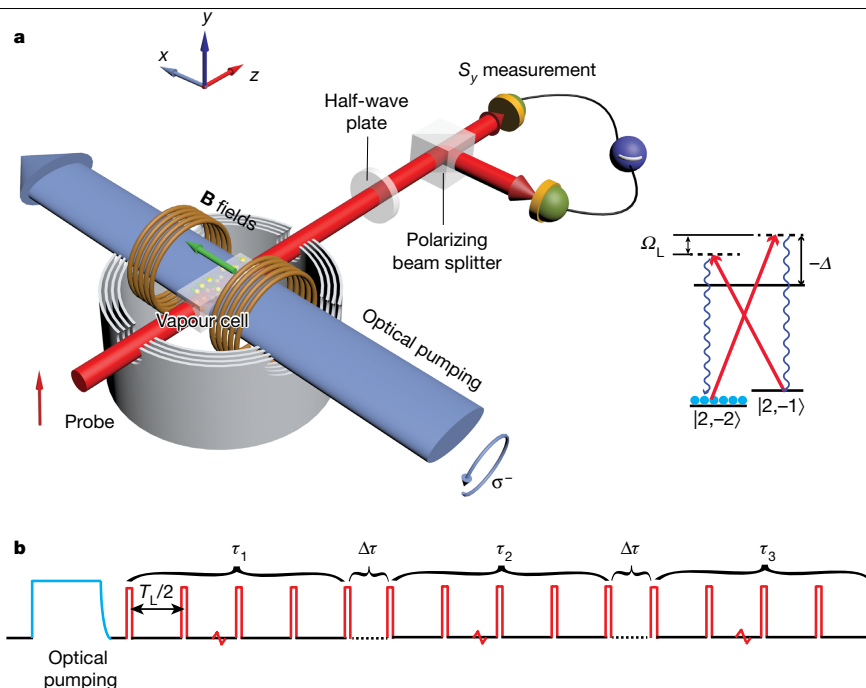


Fig. 1 | Experimental setup. **a**, Schematic of the setup. A paraffin-coated $20\text{ mm} \times 7\text{ mm} \times 7\text{ mm}$ rectangular vapour cell at 53°C resides inside a four-layer magnetic shielding to screen the ambient magnetic field. The CSS is created by optical pumping, with a pump laser tuned to the Rb D1 transition $5S_{1/2}$ $F=2 \rightarrow 5P_{1/2}$ $F'=2$ and a repump laser stabilized to the Rb D2 transition $5S_{1/2}$ $F=1 \rightarrow 5P_{3/2}$ $F'=2$, both with σ^- circular polarization along the x direction. A magnetic field (along the x direction) of 0.71 G is applied to induce a ground-state Zeeman splitting (that is, a Larmor frequency of $\Omega_L \approx 2\pi \times 500\text{ kHz}$) and to hold the collective spin. A linearly polarized laser beam, which is blue-detuned by 2.1 GHz from the $5S_{1/2}$, $F=2 \rightarrow 5P_{3/2}$, $F'=3$ transition of the D2 line and propagates in the z direction, probes the quantum

fluctuations of the spin. The Stokes component S_y is measured using a balanced polarimetry scheme and detected at the Larmor frequency Ω_L by a lock-in amplifier. **b**, Pulse sequence. The pump lasers prepare the atoms in the CSS and are then turned off adiabatically (see Methods). They are followed by the stroboscopic probe pulses, which are spaced by half the Larmor period, $T_L/2$. The first part (pulse duration τ_1) of the probe, called squeezing pulse, creates entanglement between S_y and J_z . J_z is squeezed through the detection of S_y , and the second part (pulse duration τ_2), called the verification pulse, verifies the squeezing. The state is further probed (squeezed) for a duration of τ_3 . The time $\Delta\tau = 0.3\text{ ms}$ between the three probe periods is to avoid interference from the lock-in amplifier.

later measurements on the system. The incorporation of later measurements supplements the well established measurement-based entanglement generation protocol^{5–10,24} and provides further information about measurement outcomes at intermediate times. The combined information from prior and posterior measurements on the collective spin of $N_{\text{at}} = 1.87 \times 10^{11}$ hot atoms in a vapour cell is equivalent to a noise reduction of 5.6 dB and a spin squeezing of 4.5 dB using the Wineland criterion, and corresponds to an angular spin variance of $4.6 \times 10^{-13}\text{ rad}^2$. In the following, we refer to this noise reduction as ‘squeezing’, but we recall that we are referring to the squeezing of an outcome probability distribution, not of a physical state.

Consider a collective atomic spin given by the sum of the total angular momenta of individual atoms, $\hat{J}_i = \sum_k \hat{J}_i^k$, with $i = x, y, z$. The macroscopic spin orientation J_x is along the applied bias magnetic field \mathbf{B} , and the collective spin components $\hat{J}_{y,z}$ oscillate in the laboratory frame at the Larmor frequency Ω_L . In the rotating frame, they obey the commutation relation $[\hat{J}_y, \hat{J}_z] = i\hat{J}_x$ ($\hbar = 1$; \hbar , reduced Planck constant).

The QND measurement of the collective atomic spin is realized by coupling the atomic ensemble to a light beam with the off-resonant Faraday interaction described in equation (1), such that a direct measurement on the transmitted field provides information about the atomic spin^{10,25}:

$$\hat{H}_{\text{int}} = \frac{\sqrt{2}\kappa}{\sqrt{N_{\text{ph}}N_{\text{at}}}} \hat{J}_z \hat{S}_z \quad (1)$$

Here N_{ph} is the number of photons in a pulse of duration τ and N_{at} is the atom number. \hat{S}_z is the Stokes operator of the probe light, relating to

the photon number difference between σ^+ and σ^- polarization. The coupling constant $\kappa^2 \propto d_0\eta \propto N_{\text{ph}}N_{\text{at}}$ characterizes the measurement strength in QND detection, with d_0 the resonant optical depth and η the atomic depumping rate causing decay of the collective spin.

We use a ^{87}Rb ensemble of 10^{11} atoms contained in a paraffin-coated vapour cell²⁶, as shown in Fig. 1. The coating provides a spin-protecting environment, enabling high-performance optical pumping and allowing the long spin coherence time to reduce the information loss due to decoherence. The atoms are initially prepared in the state $5S_{1/2}$ $|F=2, m_F=-2\rangle$ (defined by the quantization axis x) by optical pumping propagating along the x direction parallel to the \mathbf{B} field. We achieve up to 97.9% polarization of the spins, resulting in a 6% increase of the measured variance compared to the fully polarized coherent spin state (CSS). The quantum fluctuations of the spin are probed by a linearly polarized off-resonant D2 laser beam propagating in the z direction. The projection noise limit is calibrated by measuring the noise of the collective spin of the unpolarized sample, which is 1.25 times that of the CSS (see Methods). The QND measurement of the spin component \hat{J}_z is achieved by implementing the stroboscopic quantum back-action evasion protocol¹⁰ (that is, modulating the measurement intensity at twice the Larmor frequency with an optimal duty factor of 14%).

To describe the atomic system and its collective spin fluctuations during the optical probing, we apply the general quantum theory of measurements. To account for a quantum state conditioned on both prior and posterior probing of a quantum system, we consider a system subject to three subsequent measurement processes. Each measurement (i) is described as a general positive-operator-valued

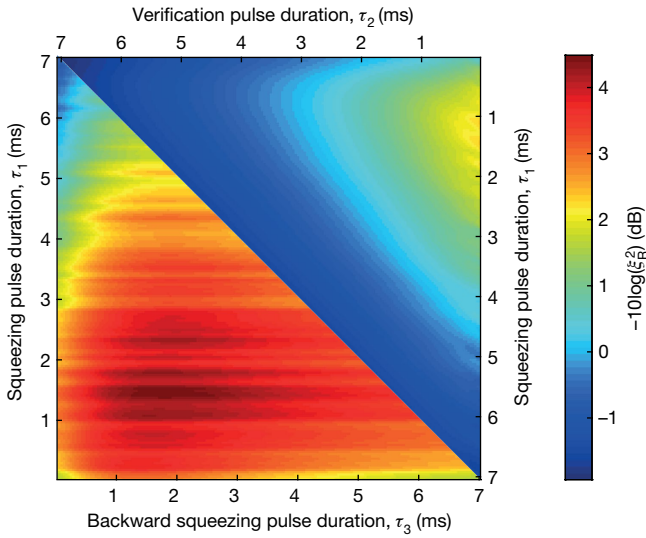


Fig. 2 | Experiment results. The lower diagonal shows the degree of spin squeezing (see colour bar) of the three-pulse scheme for various time durations of the first and third pulses. The duration of the second probe pulse is 0.037 ms. Better squeezing is observed for a shorter verification pulse τ_2 , which minimizes the disturbance of the state prepared during the first pulse. The squeezing reaches its maximum value of 4.5 dB at $\tau_1 = 1.4$ ms and $\tau_3 = 1.7$ ms, as an optimal balance between the increased atom–light coupling strength with the higher photon number and the spin decoherence due to spontaneous emission. The upper diagonal shows the spin squeezing detected when using the traditional squeezing and verification two-pulse scheme as a function of τ_1 and τ_2 . The best squeezing here is 2.3 dB. The probe laser has an average power of 1.18 mW in both experiments. ξ_R^2 is the squeezing parameter according to the Wineland criterion (see Supplementary Information).

measurement (POVM) with a set of operators $\{\hat{\Omega}_m^{(i)}\}$ associated with the measurement outcome m and fulfilling $\sum_m \hat{\Omega}_m^{(i)\dagger} \hat{\Omega}_m^{(i)} = \hat{\mathbb{I}}$, where $\hat{\mathbb{I}}$ is the identity matrix. For a system represented by the density matrix ρ at the time of a measurement, the probability of measuring outcome m is

$$\Pr^{(i)}(m) = \text{tr}(\hat{\Omega}_m^{(i)} \rho \hat{\Omega}_m^{(i)\dagger}) \quad (2)$$

and the resulting conditional state reads

$$\rho|_m = \frac{\hat{\Omega}_m^{(i)} \rho \hat{\Omega}_m^{(i)\dagger}}{\Pr^{(i)}(m)} \quad (3)$$

Assuming no further dynamics between the measurements, we can evaluate the joint probability that three subsequent measurements, described by $\{\hat{\Omega}_m^{(i)}\}$, yield outcomes m_1 , m_2 and m_3 as

$$\Pr(m_1, m_2, m_3) = \text{tr}(\hat{\Omega}_{m_3}^{(3)} \hat{\Omega}_{m_2}^{(2)} \hat{\Omega}_{m_1}^{(1)} \rho \hat{\Omega}_{m_1}^{(1)\dagger} \hat{\Omega}_{m_2}^{(2)\dagger} \hat{\Omega}_{m_3}^{(3)\dagger}) \quad (4)$$

This equation can be factored into: (i) the probability of obtaining the first outcome, m_1 , (ii) the probability of obtaining outcome m_2 in the state conditioned on the first outcome, and (iii) the probability of obtaining outcome m_3 in the state conditioned on the first two outcomes. This is equivalent to the conventional evolution of quantum trajectories, where the quantum state—and hence the probability of a measurement outcome—depends on previous measurements. However, the joint probability distribution (4) also permits evaluation of the probability of, for example, the second measurement, conditioned on the outcome of the first and the last one

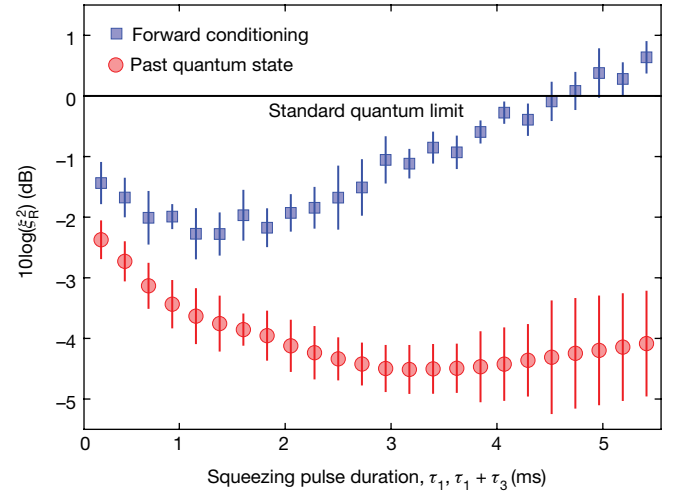


Fig. 3 | Squeezing versus total squeezing pulse duration in two- and three-pulse schemes. The horizontal axis shows τ_1 for the two-pulse scheme (forward conditioning) and $\tau_1 + \tau_3$ for the three-pulse scheme (PQS protocol). τ_2 is 0.037 ms for both curves. The attainable squeezing for the three-pulse scheme is greater and has a better long-time behaviour than the two-pulse scheme. The error bars (1 s.d.) are derived from 10 identical experiments, each consisting of 10,000 repetitions of the pulse sequence shown in Fig. 1b.

$$\Pr(m_2|m_1, m_3) = \Pr(m_1, m_2, m_3) / \sum_{m_2} \Pr(m_1, m_2', m_3) \quad (5)$$

where m_1 and m_3 are fixed to the observed values and the denominator is merely a normalization factor.

Using equation (4) and the cyclic permutation property of the trace, we can write this probability as³

$$\Pr_p(m_2, t) = \frac{\text{tr}(\hat{\Omega}_{m_2}^{(2)} \rho|_{m_1} \hat{\Omega}_{m_2}^{(2)\dagger} E|_{m_3})}{\sum_{m'} \text{tr}(\hat{\Omega}_{m'}^{(2)} \rho|_{m_1} \hat{\Omega}_{m'}^{(2)\dagger} E|_{m_3})} \quad (6)$$

where $\rho|_{m_1}$ is the state conditioned on the first measurement (see equation (3)) and $E|_{m_3} = \hat{\Omega}_{m_3}^{(3)\dagger} \hat{\Omega}_{m_3}^{(3)}$.

We observe that the conventional expression for the outcome probabilities in equation (2) depending only on the density matrix $\rho|_{m_1}$, conditioned on the prior evolution, is supplemented with the operator $E|_{m_3}$, which depends only on the later measurement outcomes. The same formalism applies to cases with continuous sequences of measurements occurring simultaneously with Hamiltonian and dissipative evolution. Examples of how the operators $\rho(t)$ and $E(t)$ evolve to time t from the initial and final time, respectively, are given in ref.³.

The specific form of the POVM operators and their action on the quantum states in our experiments can be derived explicitly in a simplified form because our system dynamics is restricted to Gaussian states. This follows from the Holstein–Primakoff transformation that maps the spin operators perpendicular to the large mean spin on the canonical position and momentum operators, $\hat{x}_A = \hat{J}_y / \sqrt{|K J_x|}$ and $\hat{p}_A = \hat{J}_z / \sqrt{|K J_x|}$. The CSS with all atoms in $|F, m_F = -F\rangle$, characterized by $\text{Var}(\hat{J}_y) = \text{Var}(\hat{J}_z) = J_x/2 = N_{\text{at}} F/2$, is equivalent to the Gaussian ground state of a harmonic oscillator, and an excitation with the ladder operator \hat{b}^\dagger corresponds to a quantum of excitation distributed symmetrically among all atoms¹⁰. Similar canonical operators, $\hat{x}_L = \hat{S}_y / \sqrt{|K S_x|}$ and $\hat{p}_L = \hat{S}_z / \sqrt{|K S_x|}$, and Gaussian states describe the probe field degrees of freedom (see Supplementary Information).

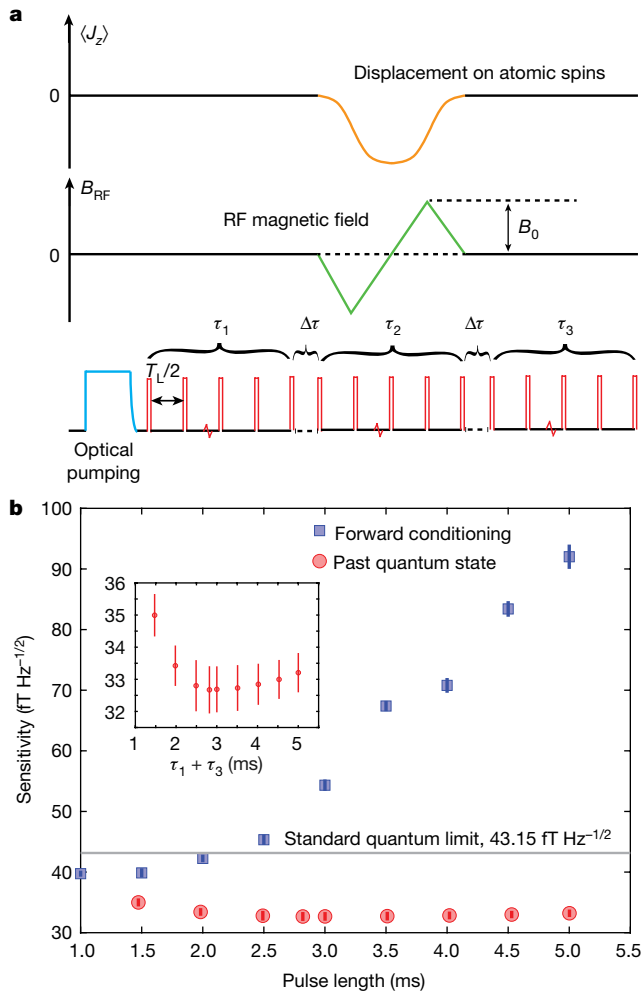


Fig. 4 | PQS-enhanced magnetometry. **a**, Pulse sequence. An RF magnetic field pulse oscillating at the Larmor frequency is switched on during the second probe sequence τ_2 in the direction orthogonal to the static \mathbf{B} field. The amplitude of the RF field, B_{RF} , is modulated as a zero-area two-triangle profile. **b**, Sensitivity of the two- and three-pulse schemes as a function of the duration of the squeezing pulses. The horizontal axis shows τ_1 for the two-pulse scheme (forward conditioning) and $\tau_1 + \tau_3$ for the three-pulse scheme (PQS protocol). $\tau_2 = 1$ ms for both curves. Similar to the squeezing results in Fig. 3, the sensitivity of the three-pulse scheme is superior to that of the two-pulse scheme and has a better long-time behaviour. The error bars (1 s.d.) are derived from five identical experiments, each consisting of 2,000 repetitions. The grey line represents the sensitivity imposed by the standard quantum limit in our system. The inset magnifies the sensitivity scale for the PQS results.

In Supplementary Information we show that the measurement operator $\hat{\Omega}_m$ in equation (2) acting on the atomic state upon readout of the value m of the field quadrature \hat{x}_L is given by $\hat{\Omega}_m = \int \psi_{\hat{x}_L}(m - ka) |a\rangle \langle a|_{\hat{p}_A} da$, where $\psi_{\hat{x}_L}(m) = \frac{1}{\pi^{1/4}} \exp(-\frac{m^2}{2})$ characterizes the quadrature distribution of the input coherent state of the probe laser beam.

For two successive QND measurements with coupling strengths κ_1 and κ_2 , the POVM formalism shows that the second outcome is governed by a Gaussian distribution with a mean value conditioned on the first outcome (see Supplementary Information)

$$\Pr(m_2|m_1) = \frac{1}{\sqrt{\pi}\sigma} \exp \left[-\frac{\left(m_2 - \frac{\kappa_2 m_1 \kappa_1}{1 + \kappa_1^2} \right)^2}{2\sigma^2} \right] \quad (7)$$

Here the variance $\sigma^2 = \frac{1}{2} + \frac{1}{2} \frac{\kappa_2^2}{1 + \kappa_1^2}$ is composed of a contribution of 1/2 from the light shot noise and a contribution from the atomic spin, which is reduced by the conditional spin squeezing by the first measurement with strength κ_1 .

If the spin oscillator is further probed by a third QND pulse with coupling strength κ_3 and measurement outcome m_3 , the conditional probability for the outcome of the middle measurement is obtained as

$$\Pr(m_2|m_1, m_3) = \frac{1}{\sqrt{\pi}\sigma_p} \exp \left\{ -\frac{\left[m_2 - \frac{\kappa_2(m_1 \kappa_1 + m_3 \kappa_3)}{1 + \kappa_1^2 + \kappa_3^2} \right]^2}{2\sigma_p^2} \right\} \quad (8)$$

The past probability yields a Gaussian distribution with variance $\sigma_p^2 = \frac{1}{2} + \frac{1}{2} \frac{\kappa_2^2}{1 + \kappa_1^2 + \kappa_3^2}$. The reduction by $1 + \kappa_1^2 + \kappa_3^2$ shows that the incorporation of the information from later measurements has a similar effect as increasing the coupling strength of the first probing from κ_1^2 to $\kappa_1^2 + \kappa_3^2$.

Experimentally, for the normal two-pulse scheme of forward-conditioning QND, we achieve the best spin squeezing of 2.3 ± 0.2 dB (Fig. 2, upper diagonal) according to the Wineland criterion²⁷ for $\tau_1 = 1.23$ ms and a conditional noise reduction of about 4.3 dB, in good agreement with the theoretical prediction (see Supplementary Information). In stark contrast, as predicted by equation (8), for the three-pulse scheme that extracts the full information from the full measurement record using the PQS, we observe an improved conditional noise reduction of about 5.6 dB and spin squeezing of 4.5 ± 0.40 dB (Fig. 2, lower diagonal) according to the Wineland criterion for $\tau_1 = 1.4$ ms and $\tau_3 = 1.7$ ms.

The main reason that the probing before and after the verification pulse sequence yields stronger squeezing than an initial longer probing sequence is the decoherence of the spins. First, owing to decoherence, the spin squeezing is gradually lost, and measurement results obtained during the early stages of the squeezing (first) pulse sequence will be less correlated with the spin ensemble at the time of the verification (second) pulse. If we instead postpone these measurements to occur in the third pulse sequence immediately after the verification pulse, the correlations will be stronger, that is, the conditional variance will be lower. Secondly, the large average spin component J_x is reduced during probing, weakening the squeezing according to the Wineland criterion. With retrodicted squeezing, the spin variance is measured relative to the mean spin at the time of the verification pulse, which has not yet suffered the reduction due to the third pulse sequence.

As shown in Fig. 3, even if we keep the total duration of the squeezing equal for both schemes, the squeezing that is attainable when using the information obtained both before and after the second pulse is better than that achieved when using only the information before the second pulse.

Although retrodiction is not a state preparation method for spin squeezing, it provides metrological advantage, as demonstrated by radio-frequency (RF) magnetometry (Fig. 4). The pulse sequence is the same as that shown in Fig. 1, but a magnetic field pulse is applied during the second pulse τ_2 to generate a temporary offset of the spin component J_z . For our proof-of-principle demonstration, this field oscillates at the Larmor frequency and follows a time-varying profile with a known shape but unknown amplitude. The procedure is outlined in Methods and summarized as follows: the value of the atomic observable p_A is retrodicted in each experiment to a certain conditional mean value and a definite variance before and after the applied magnetic field. The m_2 readout signal thus reports directly a noisy estimate of the applied field pulse, as demonstrated by the results presented in Fig. 4b. We find that the PQS protocol gives a better sensitivity than the forward conditioning protocol for the same total duration of the full pulse sequence. Notably, as expected, the sensitivity of the

three-pulse scheme experiences no substantial influence of the spin decoherence that occurs during the last detection pulse. Given $\tau_2 = 1$ ms, the best sensitivity achieved via the PQS protocol is $B_{\text{RF}}\sqrt{\tau_2}/\text{SNR} = 32.67 \pm 0.73 \text{ fT Hz}^{-1/2}$, where the signal-to-noise ratio SNR is the ratio of the mean to the standard deviation of the data obtained for $B_0 \approx 1$ pT (Fig. 4a) applied during τ_2 (ref. ¹⁹). We note that our analysis is simplified here by the QND character of the probing, whereas applications in which the non-unitary measurement back-action is interspersed with unitary rotation of the spin ellipse^{28,29} can also be handled by the more complete PQS analysis with Gaussian states⁴.

This work introduces a higher limit on the size (in terms of the number of spins) that a physical system can have while still being subjected to measurements at the quantum limit. Further improvement of the squeezing is possible by realizing a multiple light-pass scheme^{30,31} to enhance the coupling strength and incorporate unconditional spin squeezing. Atoms constitute ideal high-sensitivity probes for a number of physical phenomena^{21,22}, and our retrodiction procedure may affect the practical applications of quantum sensors. In particular, the retrodicted evolution of physical systems may offer insight and allow precision estimation of time-dependent perturbations³² that are applicable, for example, to force sensing with mechanical oscillators^{23,33}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2243-7>.

- Giovannetti, V., Lloyd, S. & Maccone, L. Advances in quantum metrology. *Nat. Photon.* **5**, 222–229 (2011).
- Wiseman, H. M. & Milburn, G. *Quantum Measurement and Control* (Cambridge Univ. Press, 2010).
- Gammelmark, S., Julsgaard, B. & Mølmer, K. Past quantum states of a monitored system. *Phys. Rev. Lett.* **111**, 160401 (2013).
- Zhang, J. & Mølmer, K. Prediction and retrodiction with continuously monitored Gaussian states. *Phys. Rev. A* **96**, 062131 (2017).
- Hosten, O., Engelsen, N. J., Krishnakumar, R. & Kasevich, M. A. Measurement noise 100 times lower than the quantum-projection limit using entangled atoms. *Nature* **529**, 505–508 (2016).
- Cox, K. C., Greve, G. P., Weiner, J. M. & Thompson, J. K. Deterministic squeezed states with collective measurements and feedback. *Phys. Rev. Lett.* **116**, 093602 (2016).
- Appel, J. et al. Mesoscopic atomic entanglement for precision measurements beyond the standard quantum limit. *Proc. Natl Acad. Sci. USA* **106**, 10960–10965 (2009).
- Schleier-Smith, M. H., Leroux, I. D. & Vuletić, V. States of an ensemble of two-level atoms with reduced quantum uncertainty. *Phys. Rev. Lett.* **104**, 073604 (2010).
- Chaudhury, S. et al. Quantum control of the hyperfine spin of a Cs atom ensemble. *Phys. Rev. Lett.* **99**, 163002 (2007).
- Vasilakis, G. et al. Generation of a squeezed state of an oscillator by stroboscopic back-action-evading measurement. *Nat. Phys.* **11**, 389–392 (2015).
- Mølmer, K. & Madsen, L. B. Estimation of a classical parameter with Gaussian probes: magnetometry with collective atomic spins. *Phys. Rev. A* **70**, 052102 (2004).
- Aharonov, Y., Albert, D. Z. & Vaidman, L. How the result of a measurement of a component of the spin of a spin-1/2 particle can turn out to be 100. *Phys. Rev. Lett.* **60**, 1351 (1988).
- Aharonov, Y. & Vaidman, L. Properties of a quantum system during the time interval between two measurements. *Phys. Rev. A* **41**, 11–20 (1990).
- Aharonov, Y. & Vaidman, L. Complete description of a quantum system at a given time. *J. Phys. A* **24**, 2315–2328 (1991).
- Rybarczyk, T. et al. Forward-backward analysis of the photon-number evolution in a cavity. *Phys. Rev. A* **91**, 062116 (2015).
- Tan, D., Weber, S. J., Siddiqi, I., Mølmer, K. & Murch, K. W. Prediction and retrodiction for a continuously monitored superconducting qubit. *Phys. Rev. Lett.* **114**, 090403 (2015).
- Rossi, M., Mason, D., Chen, J. & Schliesser, A. Observing and verifying the quantum trajectory of a mechanical resonator. *Phys. Rev. Lett.* **123**, 163601 (2019).
- Shah, V., Vasilakis, G. & Romalis, M. V. High bandwidth atomic magnetometry with continuous quantum nondemolition measurements. *Phys. Rev. Lett.* **104**, 013601 (2010).
- Wasilewski, W. et al. Quantum noise limited and entanglement-assisted magnetometry. *Phys. Rev. Lett.* **104**, 133601 (2010).
- Martin Ciurana, F., Colangelo, G., Slodička, L., Sewell, R. J. & Mitchell, M. W. Entanglement-enhanced radio-frequency field detection and waveform sensing. *Phys. Rev. Lett.* **119**, 043603 (2017).
- Smiciklas, M., Brown, J. M., Cheuk, L. W., Smullin, S. J. & Romalis, M. V. New test of local Lorentz invariance using a ²³Ne-Rb-K comagnetometer. *Phys. Rev. Lett.* **107**, 171604 (2011).
- Bear, D., Stoner, R. E., Walsworth, R. L., Kostelevy, V. A. & Lane, C. D. Limit on Lorentz and CPT violation of the neutron using a two-species noble-gas maser. *Phys. Rev. Lett.* **85**, 5038 (2000).
- Khalili, F. Ya. & Polzik, E. S. Overcoming the standard quantum limit in gravitational wave detectors using spin systems with a negative effective mass. *Phys. Rev. Lett.* **121**, 031101 (2018).
- Kong, J., Jiménez-Martínez, J., Troullinou, C., Lucivero, V. G. & Mitchell, M. W. Measurement-induced nonlocal entanglement in a hot, strongly-interacting atomic system. Preprint at <http://arXiv.org/quant-ph/1804.07818> (2018).
- Hammerer, K., Sørensen, A. S. & Polzik, E. S. Quantum interface between light and atomic ensembles. *Rev. Mod. Phys.* **82**, 1041–1093 (2010).
- Balabas, M. V., Karaulanov, T., Ledbetter, M. P. & Budker, D. Polarized alkali-metal vapor with minute-long transverse spin-relaxation time. *Phys. Rev. Lett.* **105**, 070801 (2010).
- Wineland, D. J., Bollinger, J. J., Itano, W. M. & Heinzen, D. J. Squeezed atomic states and projection noise in spectroscopy. *Phys. Rev. A* **50**, 67 (1994).
- Borregaard, J. & Sørensen, A. S. Near-Heisenberg-limited atomic clocks in the presence of decoherence. *Phys. Rev. Lett.* **111**, 090801 (2013).
- Braverman, B. et al. Near-unitary spin squeezing in Yb-171. *Phys. Rev. Lett.* **122**, 223203 (2019).
- Wang, M. F. et al. Two-axis-twisting spin squeezing by multipass quantum erasure. *Phys. Rev. A* **96**, 013823 (2017).
- Takeuchi, M. et al. Spin squeezing via one-axis twisting with coherent light. *Phys. Rev. Lett.* **94**, 023003 (2005).
- Tsang, M. Time-symmetric quantum theory of smoothing. *Phys. Rev. Lett.* **102**, 250403 (2009).
- Aspelmeyer, M., Kippenberg, T. J. & Marquardt, F. Cavity optomechanics. *Rev. Mod. Phys.* **86**, 1391–1452 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Experimental setup and characterization

From a technical perspective, the experimental realization of large-scale spin squeezing is challenging because classical noise amplitudes typically scale as the atom number N_{at} and dominate over that of the atom projection noise that is proportional to $\sqrt{N_{\text{at}}}$. Also, for large atomic ensembles it is difficult to achieve a uniform atom–light coupling across the entire ensemble, which is required for state preparation, manipulation and detection.

Meanwhile, strict orthogonality is required between the polarized spin and the wave vector of the probe field to avoid influence of the large polarized spin component in the y – z plane on the quantum noise measurement. In the alignment optimization, we used the intensity-modulated pump field as in a Bell–Bloom magnetometer configuration³⁴, and we found that an adiabatic turn-off of the pump pulse was necessary to minimize classical noise (see below).

Preparation and characterization of the atomic state. A d.c. magnetic field in the x direction creates the Zeeman splitting. Circularly polarized optical pumping and repumping beams along the x direction prepare the highly oriented spin states, which is crucial for the interface between light and atoms. As shown in the inset of Extended Data Fig. 1b, the pump and repump lasers are tuned to the ⁸⁷Rb D1 and D2 transitions, respectively.

Number of atoms in the vapour cell. To determine the number of atoms in the vapour cell, a Faraday rotation measurement is employed. A linearly polarized probe light travels through the atomic ensemble in the x direction. The almost fully oriented spins along the probe propagation direction cause the polarization of the probe light to rotate with Faraday angle θ . Assuming that the ensemble is fully polarized ($j_x = 2$), the number of atoms N_{at} can be estimated from θ by³⁵

$$N_{\text{at}} = -\frac{32\pi\theta V\Delta}{a_1(\Delta)\Gamma\lambda^2 j_x l_c} \quad (9)$$

where l_c is the path length in the x direction, V is the volume of the cell, $\lambda = 780$ nm is the wavelength of the probe light and $\Gamma = 2\pi \times 6.067$ MHz is the full-width at half-maximum linewidth of the excited state. The vector polarizability a_1 is given in Supplementary Information as a function of the laser detuning Δ .

Atomic population. We use the magneto-optical resonance signal (MORS) method to characterize the atomic polarization³⁶. In the experiment, a d.c. magnetic field B_x induces the Larmor precession at $\Omega_L = g_F \mu_B B_x / \hbar$ and a quadratic Zeeman splitting. A short RF magnetic field pulse at frequency Ω_L along the z direction is applied at the end of the optical pumping pulse to excite a $\Delta m = 1$ coherence between the magnetic sublevels. The subsequent spin evolution is measured through Faraday interaction with a weak linearly polarized probe beam propagating in the z direction. In Extended Data Fig. 1, the spin evolution after the short RF pulse and the corresponding Fourier transformation are plotted. The spin polarization is estimated to be 97.9% by fitting the experimental data to the model of ref.³⁶.

As a result of the imperfect optical pumping, a small fraction of the atoms remain in the $F = 1$ manifold. This amount can also be estimated using the MORS method, with the laser tuned close to the D2 $F = 1 \rightarrow F'$ transitions. The RF pulse excites $\Delta m = 1$ coherences in the $F = 2$ and $F = 1$ manifolds. The frequency of the $\Delta m = 1$ coherence for $F = 1$ is about 0.4% higher than that for $F = 2$, so we can distinguish them in the frequency domain, and we estimate the population in the $F = 1$ manifold to be less than 5% under the application of optical pumping, causing negligible effects in noise calibration.

The effective coupling strength $\tilde{\kappa}^2$ is calibrated by measuring the spin noise of the unpolarized atomic ensemble with equal population on all $F = 1$ and $F = 2$ ground states. The measured spin noise of the unpolarized sample is 1.25 times that of the CSS for the following reasons. The atoms in both the unpolarized state and the CSS are uncorrelated, so

$$\text{Var}(j_z) = \sum_{i=1}^{N_{\text{at}}} \text{Var}(j_z^i) \quad (10)$$

In the CSS, $\text{Var}(j_z) = \text{Var}(j_y) = \frac{j_x^2}{2} = 1$, whereas in the unpolarized state the spin is symmetric, which means $\text{Var}(j_z) = \langle j_z^2 \rangle = \langle j_x^2 \rangle = \langle j_y^2 \rangle = \frac{F(F+1)}{3} = 2$ for $F = 2$. When all sublevels, including three $F = 1$ states that are not observed in the measurement, have the same population, the contribution of the five $F = 2$ sublevels to the observed noise is $2 \times \frac{5}{8} N_{\text{at}} = \frac{5}{4} N_{\text{at}}$. Whereas for the CSS, the observed noise should be $1 \times N_{\text{at}}$.

In our experiment, we use light to measure the spin noise. Thus, the total noise includes the light shot noise and spin noise. So we have

$$\tilde{\kappa}^2 = \left[\frac{\text{Var}(\hat{S}_y^{\text{thermal}})}{\text{Var}(\hat{S}_y^{\text{light}})} - 1 \right] \times 0.8 \quad (11)$$

Here $\hat{S}_y^{\text{thermal}}$ and \hat{S}_y^{light} are the Stokes components acquired when measuring the unpolarized spin noise and photon shot noise, respectively.

When measuring the photon shot noise, the Larmor frequency is tuned far away from the lock-in detection bandwidth by changing the d.c. magnetic field, ruling out the noise contribution from spin noise. Extended Data Fig. 2 shows the dependence of photon shot noise on the input probe power, and the linearity demonstrates the behaviour of the photon shot noise limit, because for the coherent state of light the variances of \hat{S}_y and \hat{S}_z should satisfy $\text{Var}(\hat{S}_y) = \text{Var}(\hat{S}_z) = \frac{S_x}{2}$.

QND character of the measurement. In Extended Data Fig. 3, the coupling strength $\tilde{\kappa}^2$ and the atomic noise variance in the state prepared by optical pumping are plotted as functions of the atomic number. The observed linear scaling of spin noise power indicates a quantum limited performance and the QND character of the measurement. The atom number is independently measured by the off-resonant Faraday rotation, which gives an optical depth of about 70 at the operation temperature of 53.5 °C. This temperature was chosen as a trade-off to maximize the size of the atomic ensemble, prevent degradation of the paraffin coating, reduce the spin exchange process at higher temperature and attain high spin orientation.

Adiabatic turn-off of the pump fields. Even after fine-tuning the alignment of the optical pumping beams with the magnetic field, a small residual π -polarization component persists when viewing in the x -quantization basis, which, together with the σ^- component, creates unwanted ground-state coherence (associated with a superposition state $|F = 2, m_F = -2\rangle + \varepsilon |F = 2, m_F = -1\rangle$ where $\varepsilon \ll 1$) via two-photon processes, creating additional classical spin components $j_{y,z}$. Furthermore, an abrupt turn-off of the pump fields can excite more coherence owing to its broader Fourier spectrum. However, such unwanted coherence can be eliminated by slowly turning off the pump lasers as the parasitic superposition state adiabatically evolves to $|F = 2, m_F = -2\rangle$.

PQS-enhanced magnetometry

In this section we outline how the collective spin squeezing and the retrodicted spin uncertainty may benefit practical precision measurements. We consider the application of a time-dependent RF magnetic field with a slowly varying envelope of the form $B_{\text{RF}} = B_0 f(t)$, which causes a temporary displacement of the spin observable

$\langle J_z \rangle \propto F(t) \equiv \int^t f(\tau) d\tau$. We assume that the shape of $f(t)$ is known and that it is completed with $F(t) = 0$ before the last measurements so that the subsequent m_3 measurements carry no information about B_0 .

The perturbation coincides in time with the m_2 probe sequence, which hence yields a record of data proportional to the time-dependent offset of the spin $\propto F(t)B_0$ shown in the upper panel of Fig. 4a. The i th coherent probe pulse undergoes a coherent displacement by $\tilde{\kappa}_2[\hat{p}_A + F(t_i)B_0]$, and by subtracting the expectation value $\tilde{\kappa}_2\langle \hat{p}_A \rangle$ of the unperturbed atomic spin, which is inferred from the density matrix or PQS conditioned by the measured signal, we obtain a noisy estimate of $\tilde{\kappa}_2 F(t_i)B_0$. All field measurements are subject to a Gaussian error with a variance comprised of the measurement photon shot noise and $\tilde{\kappa}_2^2$ times the variance of \hat{p}_A . We have verified that the m_1 measurements, alone and in conjunction with the m_3 measurements, yield the mean value and variance of the m_2 measurements of \hat{p}_A according to equations (7) and (8). These equations thus constitute the basis for estimating the RF magnetic field amplitude B_0 . For simplicity, we disregard the measurement back-action of the individual (weak) m_2 pulses and hence treat their combined effect as an effective QND measurement of \hat{p}_A , including a time-weighted (equal weighting, for simplicity) integral of the displacement $F(t_i)B_0$. Subtracting in each run of the experiment the conditional mean spin given by equation (7) or (8) thus provides an estimate of B_0 . The uncertainty in the B_0 measurement (determining the magnetometer sensitivity) is composed of the shot noise contributions and the spin variance, σ^2 , given by equation (7) or (8). It is clear that the measurement uncertainty is reduced when we apply the PQS results, where the spin variance takes the smallest value.

Retrodiction is thus beneficial when measuring an RF magnetic field with zero mean amplitude. This inspires echo-type experiments in which, for example, B_{RF} is stable and lasts for τ_2 , but at time $\tau_2/2$ one applies a very short π pulse so the displacement caused by B_{RF} is reversed and the final displacement is zero. Similar to our experimental study, using a third probe pulse for retrodiction will improve the measurement of B_{RF} . Other time-dependent signals, including noisy signals with known governing statistics, may be inferred from the more elaborate time-dependent PQS theory, which may hence apply to many naturally occurring physical situations³⁷.

In addition, we note that the length of τ_2 is a trade-off between two factors: on the one hand, increasing τ_2 will enhance sensitivity; on the other hand, when τ_2 is comparable to the entanglement lifetime of ~ 1 ms, the conditioning protocol (both forward and especially backward) does not help. In other words, our protocols are good for measurements of relatively fast profile changes (of the RF amplitude) owing to the finite entanglement lifetime. This is also the case for other squeezing-enhanced metrology applications^{38–41}.

RF magnetic field detection and calibration. In the RF atomic-optical magnetometry, a polarized spin ensemble is prepared by optical pumping in the presence of a static magnetic field, which determines the atomic Larmor frequency. A transverse RF magnetic field $B_{RF}e^{i\Omega_L t}$ at the Larmor frequency causes the spin ensemble to precess and the angle of precession is proportional to the RF magnetic field. The spin dynamics are monitored with a weak off-resonant linearly polarized probe beam. As the probe beam travels through the atomic vapour, its plane of polarization rotates by an angle proportional to the spin component along the propagation direction according to the Faraday effect.

The Stokes component \hat{S}_y carrying the transverse spin information can be measured in a balanced polarimetry scheme in the $\pm 45^\circ$ basis. The signal at the Larmor frequency $\hat{S}_{y,c}$ is extracted³⁵ with a lock-in amplifier (Zurich Instrument). Here the subscript 'c' indicates 'cosine', the in-phase quadrature of the lock-in amplifier output. The sensitivity to the RF magnetic field is given by^{42,43} $B_{sen} = B_{min}\sqrt{T}$ (where T is the measurement time) with the minimal detectable field $B_{min} = B_{RF}/\text{SNR}$.

In practice, the signal-to-noise ratio SNR in our magnetometer is defined as

$$\text{SNR} = \frac{|\langle \hat{S}_{y,c} \rangle|}{\sqrt{\text{Var}(\hat{S}_{y,c})}} \quad (12)$$

Experimentally, a pair of Helmholtz coils oriented along the z axis generates a RF magnetic field along the z axis, perpendicular to the main spin along the x direction. The pulse sequence employed in our PQS-enhanced magnetometry is schematically shown in Fig. 4a.

In the protocol of PQS-enhanced magnetometry, the denominator in equation (12) is replaced by $\sqrt{\text{Var}(m_2|m_1, m_3)}$, with $\text{Var}(m_2|m_1, m_3)$ the variance of m_2 conditioned on the measurements before and after τ_2 , that is, m_1 and m_3 . Here, m_2 (that is, $\hat{S}_{y,c}$) is the sum of all the data points obtained during τ_2 in one sequence. In our demonstration RF-field measurement of a triangularly shaped RF profile, $B_0 = \max(B_{RF})$ is the height of the triangle. To compare the sensitivity with other magnetometers, we use the following definition of the aforementioned sensitivity $B_{sen} = B_0\sqrt{\tau_2}/\text{SNR}$.

To calibrate the RF coil, a pickup coil with $N_\omega = 30$ turns of copper wire and 8.35 mm diameter is employed, located at the position of the Rb cell, along the axis of the Helmholtz coils. The oscillating magnetic field creates a flux through the pickup coil that generates an electromotive force. When applying a sinusoidal magnetic field of frequency ω and amplitude B_{RF} , the current through the pickup coil can be found from measuring the voltage amplitude U_ω across the measurement resistor R_m (ref. 44). Then we have the relation between B_{RF} and U_ω

$$|B_{RF}| = \frac{|1 + Z_{coil}/R_m|U_\omega}{N_\omega A_{coil}\omega} \quad (13)$$

where A_{coil} is the cross-sectional area of the pickup coil. Its impedance is $Z_{coil} \approx i\omega L$ at frequency Ω_L with inductance $L \approx 30 \mu\text{H}$, because the resistance of the coil $R = 1.9 \Omega$ is much smaller than ωL at the frequency at which we usually operate ($2\pi \times 500 \text{ kHz}$). We use a spectrum analyser to read out the response generated in the pickup coil. The voltage is read out over the resistance $R_m = 50 \Omega$. The measured amplitude of the voltage is $U_\omega = \sqrt{2} U_{rms}$ where U_{rms} is the root-mean-square voltage. The measurement result is shown in Extended Data Fig. 4b, which indicates that the pickup coil's voltage is in good linear relation with the voltage output of the signal generator (Agilent E8257D).

Extended Data Figure 4b seems to indicate that, combined with equation (13), we may get a relation between the RF field B_{RF} (seen by the atom) and the signal generator's output. However, as shown in Extended Data Fig. 4b, this calibration can only be done for relatively large RF output from the signal generator, owing to excess electrical noises dominating the small electromotive-force voltage on the pickup coil. In practice, we applied a smaller magnetic field on our atoms, which could not be directly measured via the pickup coil. The possible solution is the following. Given that $B_{RF} \propto U_{app}$, where U_{app} is the applied voltage on the RF coil, we may extrapolate B_{RF} for the lower RF output range from the magnetic field amplitude measured in the higher RF output range with the pickup coil.

To prove that such extrapolation to the low range of RF output in Extended Data Fig. 4b is valid, we use the atoms to measure the RF field B_{RF} in this range, which however still partially overlaps with the range of Extended Data Fig. 4b. Indeed, we found that the atoms are much more sensitive than the pickup coil. For very small RF output from the signal generator, B_{RF} can be measured by the displacement of atomic spins but not by the pickup coil. Extended Data Fig. 4a presents the results of the magnetic field B_{RF} calibration performed by monitoring the displacement of atomic spin J_z . The setup is the same as that used in the magnetic field detection experiment. We vary the peak amplitude of the RF magnetic field and record the mean value of the sum

Article

of data points during the second sequence, that is, the mean value of m_2 . As illustrated in Extended Data Fig. 4a, the linearity of the mean value versus the RF peak amplitude is good, indicating the validity of the linearity of the RF signal generator's output reading U_{set} and the response of atomic spins, further enabling the extrapolation that we use, $B_{\text{RF}} \propto U_{\text{app}} \propto U_{\text{set}}$.

Based on the aforementioned observation, we can obtain the relation between the applied magnetic field and the output of the signal generator as

$$B_0(T) = 9.686 \times 10^{-8} \times 10^{P_{\text{set}}/20} \quad (14)$$

where P_{set} (in units of dB m) is the set output power reading of the RF signal generator. Through this calibration, we get the peak amplitude of the applied RF magnetic field in the magnetic field detection experiment, which is about 1.00 pT ($P_{\text{set}} = -97$ dB m on our signal generator).

Data availability

The datasets generated and analysed during this study are available from the corresponding authors upon reasonable request.

34. Bell, W. E. & Bloom, A. L. Optically driven spin precession. *Phys. Rev. Lett.* **6**, 280–281 (1961).
35. Shen, H. *Spin Squeezing and Entanglement with Room Temperature Atoms for Quantum Sensing and Communication*. PhD thesis, Univ. of Copenhagen (2015).
36. Julsgaard, B., Sherson, J., Sørensen, J. L. & Polzik, E. S. Characterizing the spin state of an atomic ensemble using the magneto-optical resonance method. *J. Opt. B* **6**, 5–14 (2004).
37. Yonezawa, H. et al. Quantum-enhanced optical-phase tracking. *Science* **337**, 1514–1517 (2012).

38. Huelga, S. F. et al. Improvement of frequency standards with quantum entanglement. *Phys. Rev. Lett.* **79**, 3865–3868 (1997).
39. André, A., Sørensen, A. S. & Lukin, M. D. Stability of atomic clocks based on entangled atoms. *Phys. Rev. Lett.* **92**, 230801 (2004).
40. Auzinsh, M. et al. Can a quantum nondemolition measurement improve the sensitivity of an atomic magnetometer? *Phys. Rev. Lett.* **93**, 173002 (2004).
41. Braverman, B., Kawasaki, A. & Vuletić, V. Impact of non-unitary spin squeezing on atomic clock performance. *New J. Phys.* **20**, 103019 (2018).
42. Budker, D. & Kimball, D. F. J. *Optical Magnetometry* (Cambridge Univ. Press, 2013).
43. Kominis, I. K., Kornack, T. W., Allred, J. C. & Romalis, M. V. A subfermotesla multichannel atomic magnetometer. *Nature* **422**, 596 (2003).
44. Jensen, K. *Quantum Information, Entanglement and Magnetometry with Macroscopic Gas Samples and Non-Classical Light*. PhD thesis, Univ. of Copenhagen (2011).

Acknowledgements We thank M. Balabas for assistance in the vapour cell fabrication and V. Vuletić for discussions. This work is supported by the National Key Research Program of China under grants 2016YFA0302000 and 2017YFA0304204, and the NNSFC under grants 61675047 and 91636107. K.M. acknowledges support from the Villum Foundation. H.S. acknowledges financial support from a UK Royal Society Newton International Fellowship (NF170876).

Author contributions K.M., H.S. and Y.X. conceived the idea. H.B., J.D., S.J., X.L., P.L., I.N., E.E.M., H.S. and Y.X. designed the experiment, performed the measurements and analysed the data together with all other authors. K.-F.Z. helped with the fabrication and characterization of vapour cells. H.B., M.W. and H.S. carried out the theoretical analysis under K.M.'s supervision. H.B., H.S., K.M. and Y.X. wrote the manuscript with contributions from all other authors. H.S. and Y.X. supervised the project.

Competing interests The authors declare no competing interests.

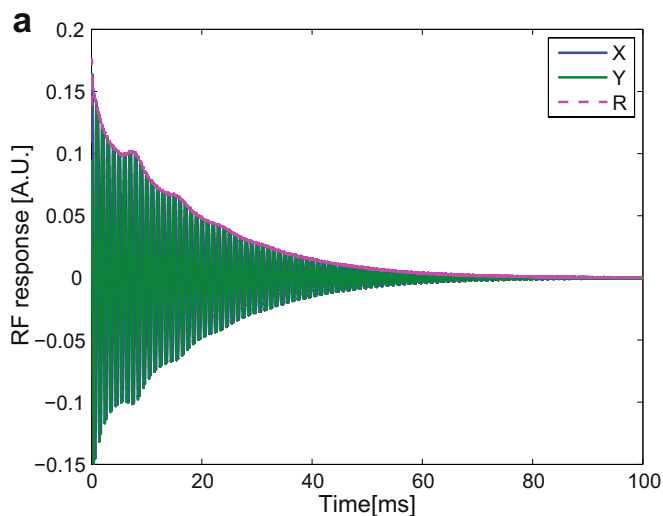
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2243-7>.

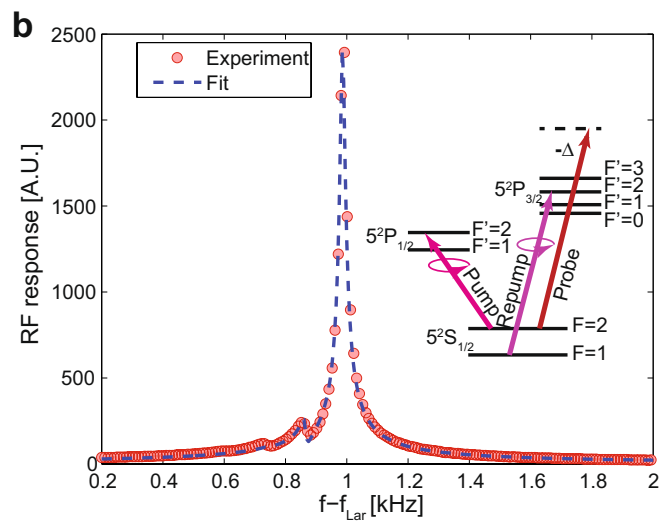
Correspondence and requests for materials should be addressed to K.M., H.S. or Y.X.

Peer review information *Nature* thanks Julian Martinez-Rincon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

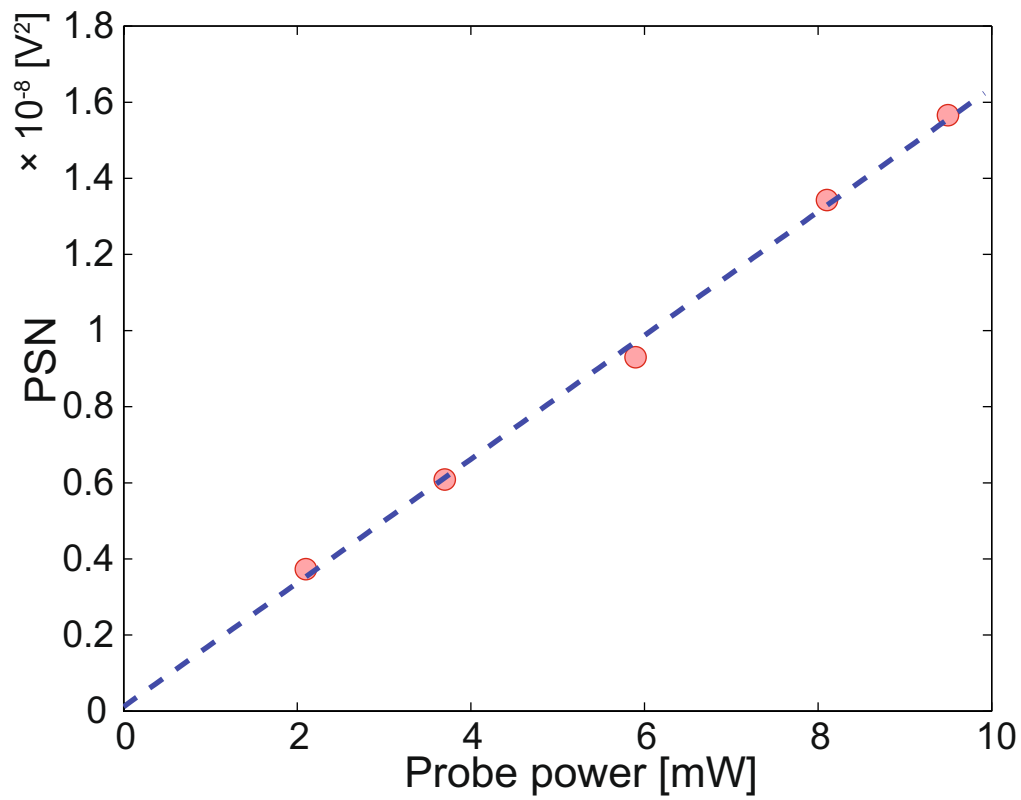
Reprints and permissions information is available at <http://www.nature.com/reprints>.



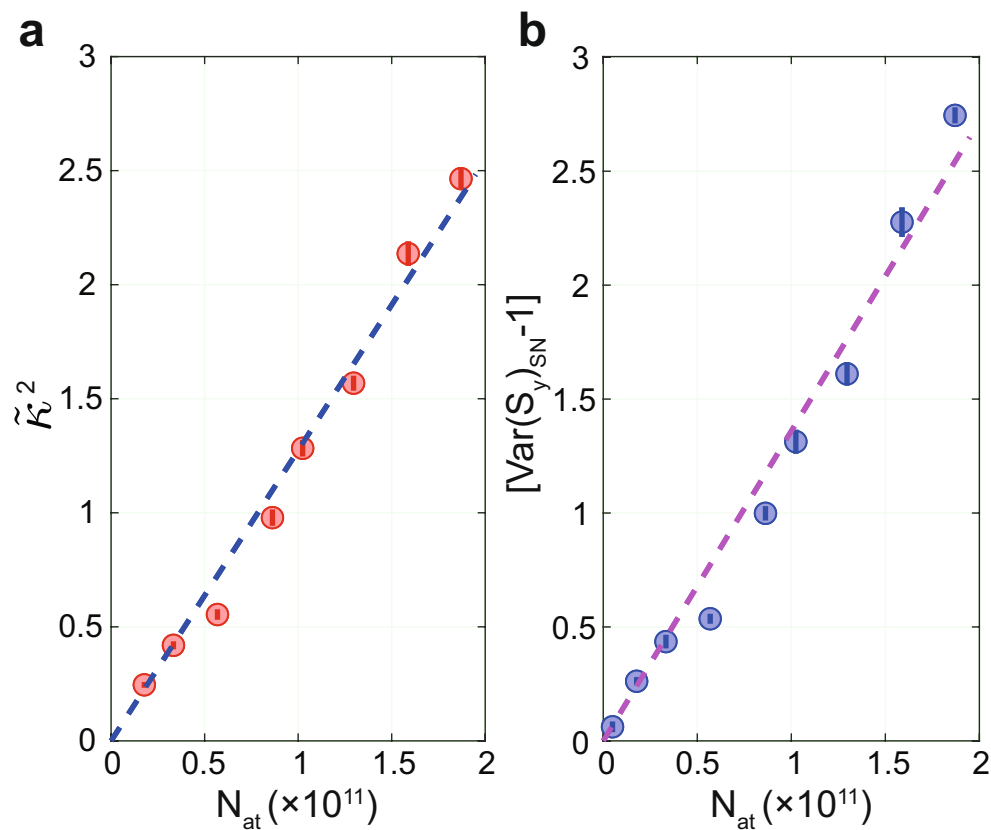
Extended Data Fig. 1 | Magneto-optical resonance signal. a. Spin response to an RF pulse. X and Y are the outputs of the lock-in amplifier, with a $\pi/2$ phase difference between them. $R = \sqrt{X^2 + Y^2}$ is the demodulated amplitude. **b.** The associated Fourier transformation of the spin response signal. f_{Lar} is the centre frequency for demodulation, with the subscript 'Lar' representing 'Larmor frequency'. f is the actual frequency of the signal before demodulation.



$f - f_{\text{Lar}}$ represents the frequency of the signal after demodulation, that is, at the lock-in amplifier output. Inset, energy levels of ^{87}Rb . All the atoms are pumped into the $F=2, m_F=-2$ state, so that they are oriented along x . The magnetic field leads to a splitting of the magnetic sublevels by the Larmor frequency Ω_L . A.U., arbitrary units.

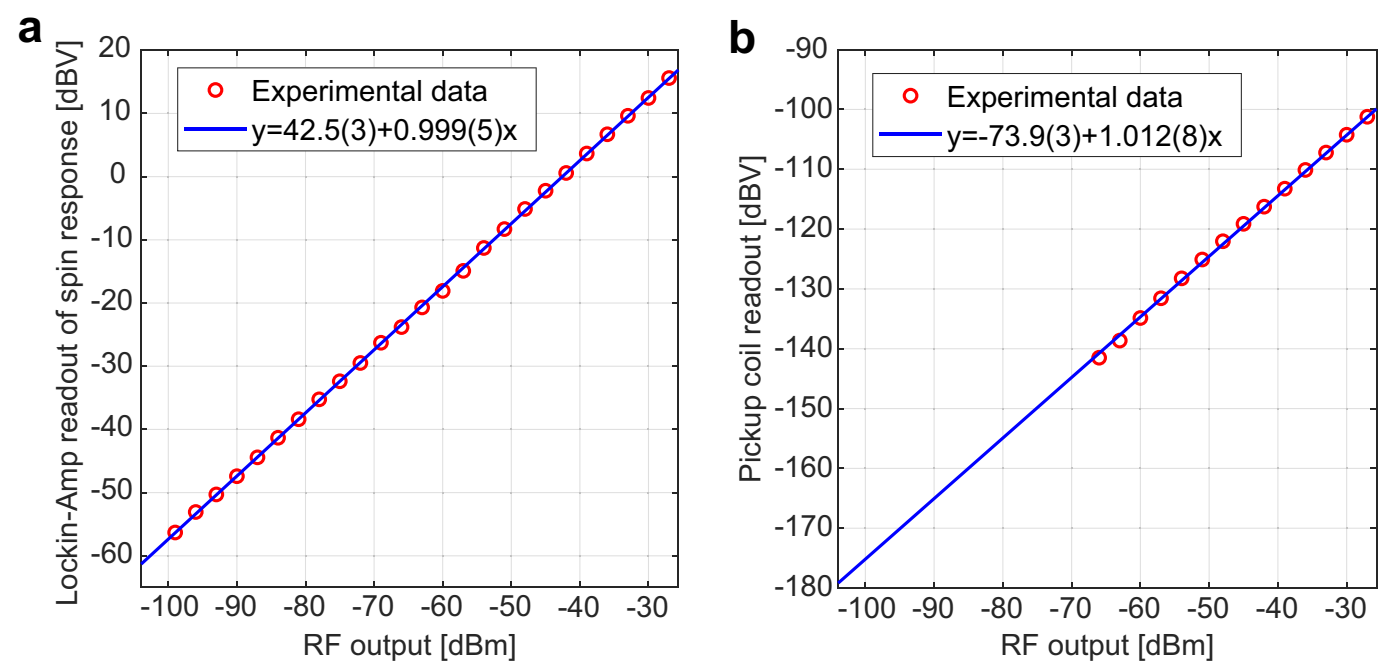


Extended Data Fig. 2 | Measured photon shot noise with different probe powers. Red circles are experimental data and the dashed line represents the linear fit of the data. PSN, photon shot noise.



Extended Data Fig. 3 | Spin noise versus atom number. **a**, Effective coupling constant $\tilde{\kappa}^2$ as a function of the number of atoms. The values of $\tilde{\kappa}^2$ are derived from the spin noise of the thermal state. **b**, Spin noise of prepared CSS versus

the number of atoms. The observed linear dependence proves that technical noise is mostly suppressed and the measured spin noise is at the projection noise limit (PNL).



Extended Data Fig. 4 | Calibration of the applied RF magnetic field.
a, Calibration using the displacement of atomic spins. **b**, Calibration using a small pickup coil. The amplitude of the RF output in our detection experiment

is -97 dB m, which lies at the bottom left of the figure. In both curves, a slope near 1 indicates a good linear relation.

Massively parallel coherent laser ranging using a soliton microcomb

<https://doi.org/10.1038/s41586-020-2239-3>

Received: 31 October 2019

Accepted: 16 March 2020

Published online: 13 May 2020

 Check for updates

Johann Riemensberger¹, Anton Lukashchuk¹, Maxim Karpov¹, Wenle Weng¹, Erwan Lucas^{1,2}, Junqiu Liu¹ & Tobias J. Kippenberg¹✉

Coherent ranging, also known as frequency-modulated continuous-wave (FMCW) laser-based light detection and ranging (lidar)¹ is used for long-range three-dimensional distance and velocimetry in autonomous driving^{2,3}. FMCW lidar maps distance to frequency^{4,5} using frequency-chirped waveforms and simultaneously measures the Doppler shift of the reflected laser light, similar to sonar or radar^{6,7} and coherent detection prevents interference from sunlight and other lidar systems. However, coherent ranging has a lower acquisition speed and requires precisely chirped⁸ and highly coherent⁵ laser sources, hindering widespread use of the lidar system and impeding parallelization, compared to modern time-of-flight ranging systems that use arrays of individual lasers. Here we demonstrate a massively parallel coherent lidar scheme using an ultra-low-loss photonic chip-based soliton microcomb⁹. By fast chirping of the pump laser in the soliton existence range¹⁰ of a microcomb with amplitudes of up to several gigahertz and a sweep rate of up to ten megahertz, a rapid frequency change occurs in the underlying carrier waveform of the soliton pulse stream, but the pulse-to-pulse repetition rate of the soliton pulse stream is retained. As a result, the chirp from a single narrow-linewidth pump laser is transferred to all spectral comb teeth of the soliton at once, thus enabling parallelism in the FMCW lidar. Using this approach we generate 30 distinct channels, demonstrating both parallel distance and velocity measurements at an equivalent rate of three megapixels per second, with the potential to improve sampling rates beyond 150 megapixels per second and to increase the image refresh rate of the FMCW lidar by up to two orders of magnitude without deterioration of eye safety. This approach, when combined with photonic phase arrays¹¹ based on nanophotonic gratings¹², provides a technological basis for compact, massively parallel and ultrahigh-frame-rate coherent lidar systems.

In recent years, interest in lidar has been fuelled by the development of autonomous driving², which requires the ability quickly to recognize and classify objects under fast-changing and low-visibility conditions¹³. Lidar can overcome the challenges of camera imaging, such as those associated with weather conditions or illumination, and has been used successfully in nearly all recent demonstrations of high-level autonomous driving¹⁴. Generally, laser ranging is based on two different principles; time-of-flight and coherent ranging¹⁵. In time-of-flight lidar, the distance of an object is determined on the basis of the delay of reflected laser pulses. To increase the speed of image acquisition, modern systems employ an array of individual lasers (as many as 256) to replace slow mechanical scanning¹⁶. The velocity information can be inferred only by comparing subsequent images, a process prone to errors caused by vehicle motion and interference.

A different principle is that of FMCW lidar^{1,4,5}. In this case a laser that is linearly chirped is sent to an object, and the time–frequency

information of the return signal is determined by delayed homodyne detection. The maximum range is therefore limited not only by the available laser power but also by the coherence length of the laser⁵. Assuming a triangular laser scan (over an excursion bandwidth B with period T ; see Fig. 1e), the distance information (that is, the time of flight, Δt) is mapped to a beat note frequency⁴, that is, $\bar{f} = \Delta t \times 2B/T$ for a static object. Owing to the relative velocity v of an object, the returning laser light is detected with a Doppler shift $\Delta f_D = \mathbf{k} \cdot \mathbf{v} / \pi$, where \mathbf{k} is the wavevector and \mathbf{v} is the velocity of the illuminated object. As a result, the homodyne return signal for a moving object is composed of two frequencies for the upwards and downwards laser scan, that is, $f_u = \bar{f} + \Delta f_D$ and $f_d = |\bar{f} - \Delta f_D|$. From the measured beat notes during one period of the scan, one can therefore determine both the distance and relative velocity of an object (see Fig. 1e). The latter greatly facilitates image processing and object classification, particularly relevant to traffic. Moreover, FMCW lidar increases the photon flux used for ranging, hence

¹Laboratory of Photonics and Quantum Measurements (LPQM), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. ²Present address: Time and Frequency Division, NIST, Boulder, CO, USA. ✉e-mail: tobias.kippenberg@epfl.ch

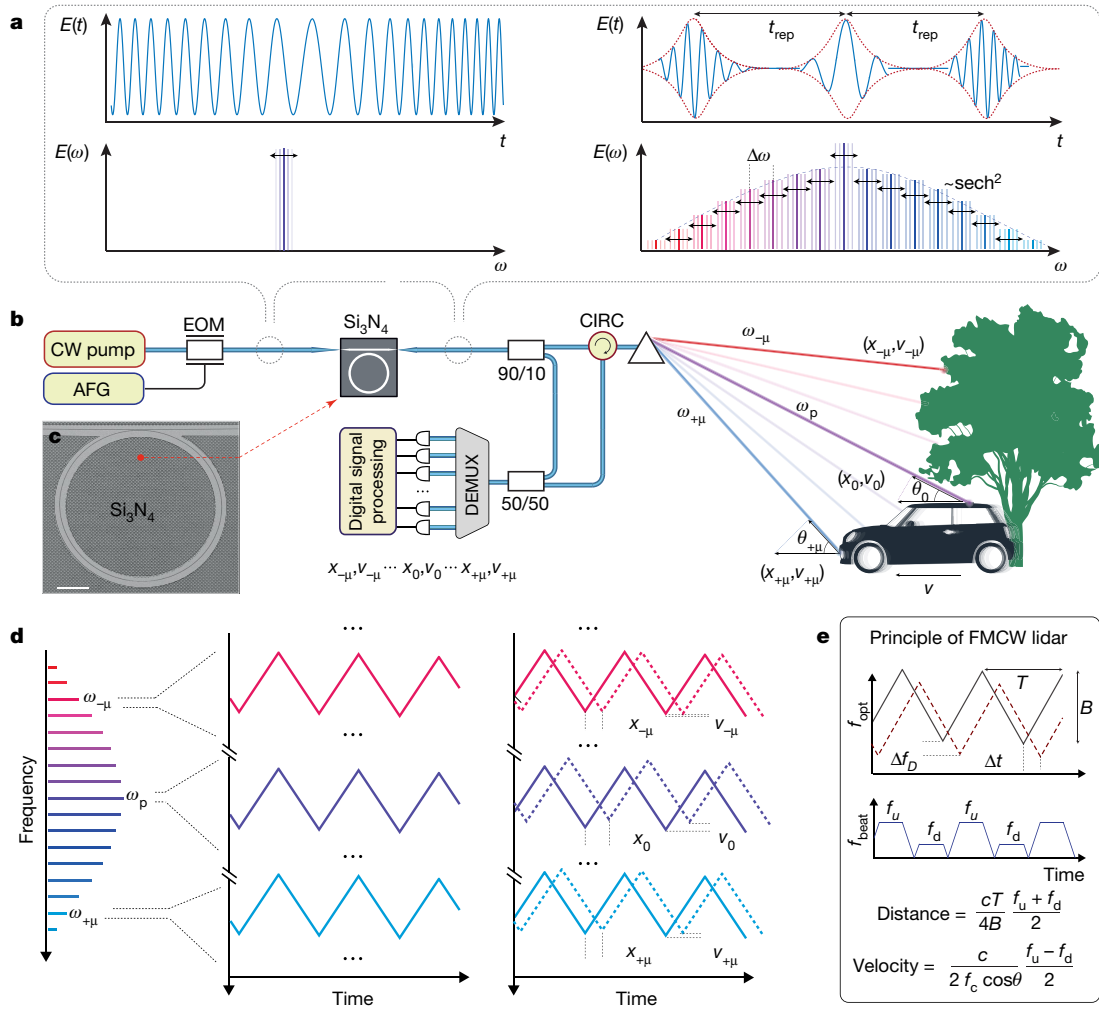


Fig. 1 | Massively parallel frequency-modulated continuous-wave lidar using soliton microcombs. **a**, Principle of DKS generation in a microresonator with a frequency agile laser. By chirping the continuous-wave (CW) pump laser, the soliton pulse stream exhibits a change in the underlying carrier, while the pulse-to-pulse repetition rate $1/t_{rep}$ and sech^2 spectral envelope remains unchanged. In the frequency domain this corresponds to scanning each individual comb tooth, that is, a change of the carrier envelope frequency f_{ceo} only. **b**, Schematic outline of the proposed system design. A frequency modulated pump laser drives a photonic integrated Si_3N_4 microresonator. Each individual sideband, spatially dispersed with diffractive optics, serves as a source of frequency-modulated laser light in a parallel detection scheme. EOM, electro-optical phase modulator; AFG, arbitrary function generator; CIRC,

optical circulator; DEMUX, demultiplexer; 50/50 and 90/10, optical splitters. **c**, Electron microscope image of 228.43- μm Si_3N_4 microring resonator. **d**, Principle of coherent velocimetry and ranging with multiple optical carriers isolated from a soliton microcomb. Interleaved upwards and downwards frequency slopes map the distance and radial velocity of target objects onto the mean and the separation of two intermediate-frequency beat tones in a delayed homodyne detection scheme. **e**, Schematic of the detected beat notes arising in coherent lidar ranging of a moving object with optical carrier frequency f_{opt} . The reflected laser light is both time-delayed and frequency-shifted owing to the Doppler effect, leading to the observation of two homodyne beat notes f_{beat} during one scan period of the laser.

increasing the sensitivity and range compared to time-of-flight lidar systems, which at present rely on sequential switching of laser diode arrays. Furthermore, coherent lidar is superior to time-of-flight implementations in low-visibility and high-background-light conditions, culminating in achievements such as detecting the distance to objects engulfed in flames¹⁷, because delayed homodyne detection makes it almost impervious to interference and malicious remote attacks¹⁸. Despite these advantages, coherent ranging suffers from the stringent requirement of narrow linewidth⁵, as well as fast and linear frequency chirping⁸, which makes massively parallel implementations, as used in time-of-flight lidar, challenging.

Concept of soliton-based parallel FMCW ranging

Here we demonstrate a massively parallel coherent FMCW source based on a soliton microcomb integrated on a photonic chip. Specifically,

we show that agile chirping of the pump laser frequency ω_p retains the soliton state and leads to simultaneous chirping of all comb teeth $\omega_{\pm\mu}$, where μ denotes the mode number comprising the soliton. The principles of massively parallel coherent lidar based on soliton microcombs are illustrated in Fig. 1a. The underlying idea is to transfer the chirp of a prepared frequency-modulated lidar source to multiple comb sidebands by using it to generate a dissipative Kerr soliton (DKS)^{19,20}. In the time domain (see Fig. 1a), we modulate the underlying soliton carrier frequency, while minimizing changes of the pulse envelope and repetition rate. In the frequency domain, this corresponds to a concurrent modulation of the optical frequency of each comb tooth around its average value (that is, a modulation of the frequency comb's carrier-envelope frequency). This effect, when combined with triangular frequency modulation of a narrow linewidth pump laser, generates a massively parallel array of independent FMCW lasers. When dispersing the channels using diffractive optics, as illustrated in Fig. 1b, each

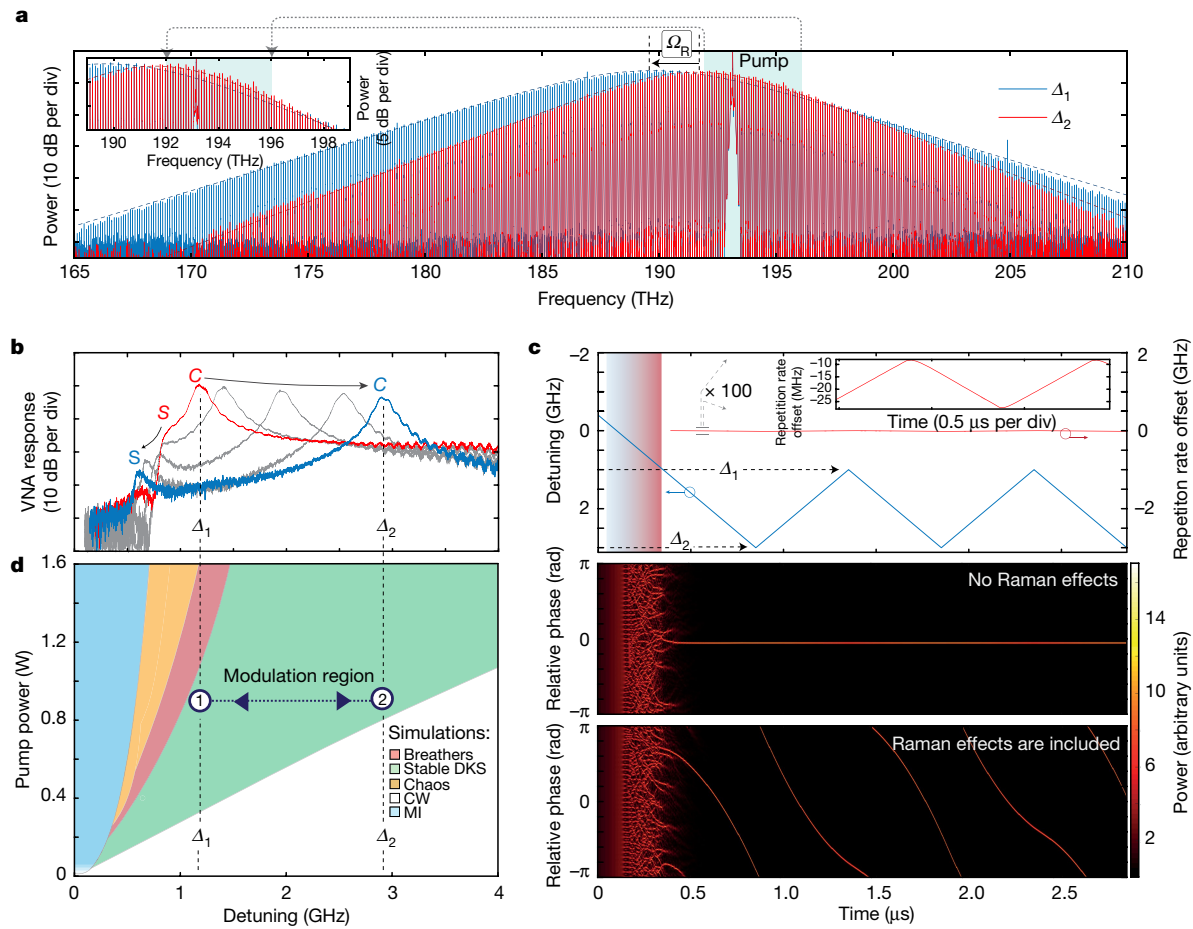


Fig. 2 | Dynamics of frequency-modulated soliton microcombs. **a**, Optical spectra of a DKS at relative laser cavity detuning $\Delta_1 = 1.2$ GHz and $\Delta_2 = 2.9$ GHz, respectively. ('div' denotes one division unit of the y axis.) The Raman self-frequency shift $\Omega_R/2\pi$ of the soliton is highlighted. The inset shows the spectral region of frequency-modulated lidar operation, showcasing individual line flatness better than 3 dB over the full pump-laser frequency excursion range. **b**, Bistable phase modulation response of DKS measured with the vector network analyser, where the S-resonance is related to the high-intensity soliton field and the C-resonance is related to the low-intensity continuous-wave field. **c**, Numerical simulation results of

frequency-modulated DKS generation. The laser is tuned through the modulation instability region (blue shaded area) and the breathing region (red shaded area) into the soliton state and triangular frequency modulation is imprinted at a chirp rate of $(d\Delta/dt) \approx 1.7 \times 10^{16} \text{ Hz}^2$. The inclusion of stimulated Raman scattering into the simulation reveals a modulation of the repetition rate of up to 10 MHz during the frequency-modulated cycle. **d**, Simulated stability chart of the soliton microcomb for the device used in the lidar experiments. The soliton existence range is highlighted in green and confined by the stability of the soliton solution. MI, modulation instability.

channel can acquire both distance and velocity information simultaneously (see Fig. 1d).

This scheme leverages three key properties of DKS; the large (that is, gigahertz) existence range of the soliton, the fact that the repetition rate changes associated with laser scanning are small, and the ability, as detailed below, to very rapidly sweep between stable operating points without destroying the soliton state or deterioration of the chirp linearity. Homodyning the reflected signal with the original comb teeth channel-by-channel, using low-bandwidth detectors and digitizers, allows the coherent ranging signal to be recovered and reconstructed for each comb line μ simultaneously, yielding velocity and distance (x_μ, v_μ) for each pixel. The scheme presented here thus enables true parallel detection of dozens and potentially hundreds of pixels simultaneously. Hence, massively parallel (and high-speed) coherent lidar becomes possible, while requiring only a single well controlled laser to generate the carrier-frequency chirped soliton. This is in contrast with dual-frequency-comb coherent time-of-flight systems^{21,22}, which achieve better distance precision and acquisition speeds, yet exhibit a limited ambiguity range dictated by the pulse repetition rate, and are challenging to parallelize because the whole frequency comb must illuminate a single pixel. In a similar fashion,

recently demonstrated coherent stitching of multiple channels from an electro-optical frequency comb generator can be used to improve the distance measurement accuracy of FMCW²³, yet demands the spectral overlap of adjacent comb modes and concurrent illumination of a single pixel.

We demonstrate here the principle of spectral multiplexing in coherent lidar employing a 99-GHz repetition rate DKS in a ultra-low-loss silicon nitride (Si_3N_4) microresonator, which is fabricated using the photonic damascene process²⁴ (see inset of Fig. 1b and Methods). Figure 2a shows the optical spectra of the DKS at the extrema of the soliton existence range with relative laser-cavity detunings $\Delta_1 = 1.2$ GHz and $\Delta_2 = 2.9$ GHz. Increasing the detuning, we observe the well known temporal compression (58 fs to 45 fs)¹⁰ and Raman self-frequency shift ($\Omega_R/2\pi = 2$ THz)²⁵ of the DKS. Interestingly, despite the frequency excursion greatly exceeding the overcoupled cavity linewidth ($\kappa_o/2\pi = 15$ MHz, $\kappa_{\text{ex}}/2\pi = 100$ MHz) the power of comb teeth between 190 THz and 200 THz does not change by more than 3 dB, thus providing more than 90 channels suitable for coherent lidar. The relative laser detuning can be inferred from the phase modulation response spectrum (see Fig. 2b), wherein the C-resonance peak of the bistable cavity field solution in the soliton existence range directly reveals the relative

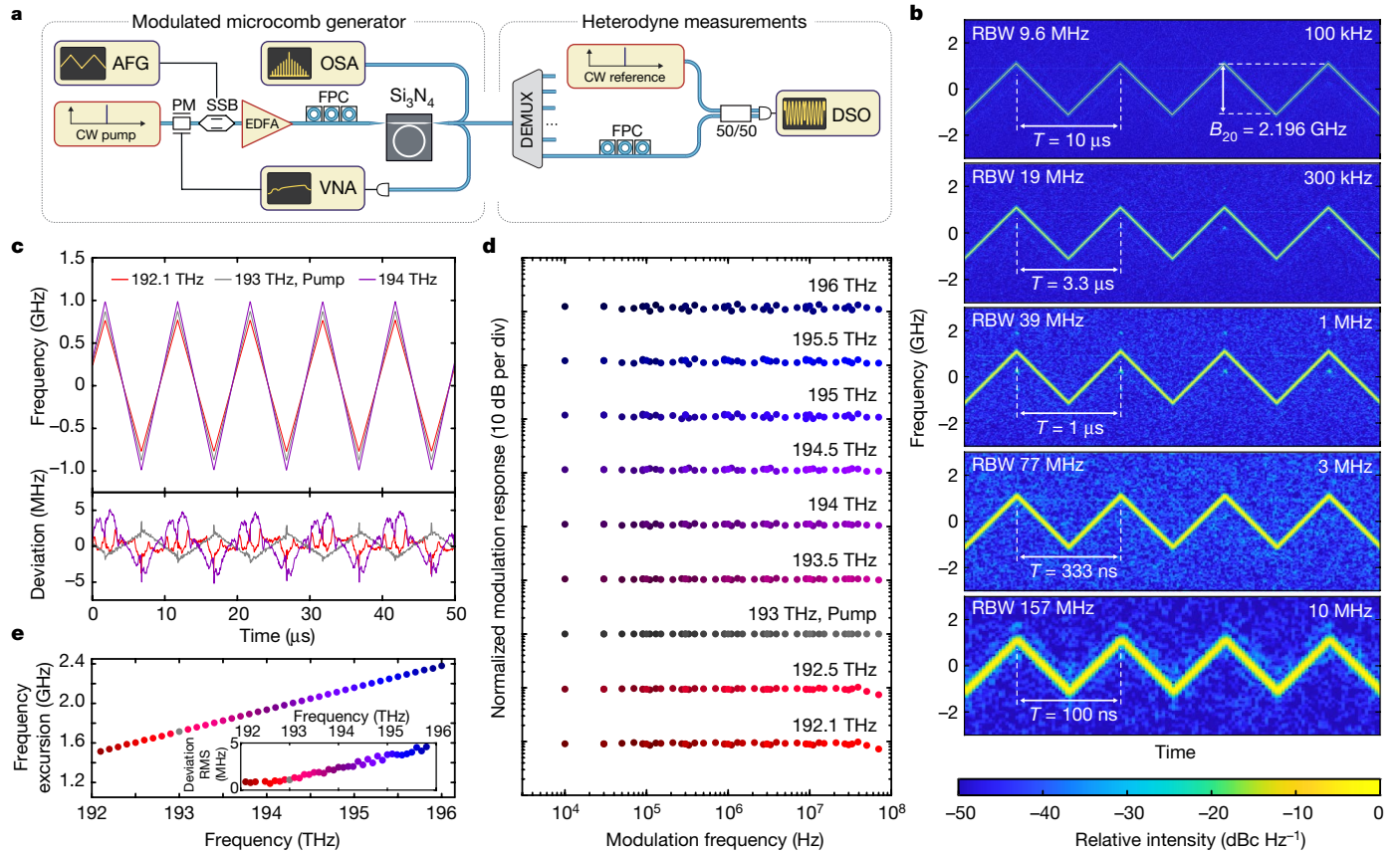


Fig. 3 | Time–frequency analysis of a chirped soliton microcomb.

a, Experimental setup. An amplified external cavity diode laser at 193 THz generates a soliton microcomb on the photonic chip. Frequency modulation is applied with a single sideband modulator (SSB). The time-dependent sideband frequencies are detected by beating with a second tuneable amplified external cavity diode laser. The optical spectrum analyser (OSA) and the vector network analyser (VNA) are used for soliton state characterization (see Fig. 2), only. DSO, Digital sampling oscilloscope; EDFA, erbium-doped fibre amplifier; FPC fibre polarization controller; PM, phase modulator; RBW, resolution bandwidth. **b**, Time–frequency maps of 1.6-GHz pump laser chirps at modulation frequencies from 10 kHz to 10 MHz, detected at the $\mu = +20$ comb sideband (195 THz); B_{20} is the frequency excursion at sideband $\mu = +20$. **c**,

Instantaneous frequency of the heterodyne beat note (top) determined by short-time Fourier transform. Deviation from a perfect triangular scan calculated by least-squares fitting (bottom) at modulation frequency 100 kHz. **d**, Pump to sideband frequency-modulated transduction determined from the frequency-modulated amplitude of the first nine harmonics of the modulation frequency of linearized frequency-modulated traces between 10 kHz (dark markers) and 10 MHz (light markers). Sideband values (colours as in **e**) are offset by 10 dB and normalized with respect to the modulation amplitudes of the pump (grey marker). **e**, Channel-dependent frequency excursion bandwidth at 100-kHz modulation frequency. The inset shows the root-mean-square (RMS) deviation from the perfectly triangular modulation pattern.

detuning between the cavity resonance and the continuous-wave pump laser²⁶.

We next perform numerical simulations based on the Lugiato–Lefever equation^{27,28}, which demonstrate the ability of the DKS state to transfer the chirp from the pump laser to all comb teeth (see Fig. 2c). The numerical laser scan is started at $\Delta = -0.4 \text{ GHz}$ and the detuning Δ is subsequently increased with a linear chirp rate of $|d\Delta/dt| = 4 \times 10^{15} \text{ Hz}^2$, tuning past the modulation instability region and exciting a single soliton. Hereafter, the linear laser scan is inflected and a symmetric triangular frequency modulation with equal chirp rate is continued. If stimulated Raman effects^{25,29} and higher-order dispersion are neglected, the repetition rate remains almost perfectly constant and the frequency chirp is faithfully transduced to each comb line. Even more surprisingly, the inclusion of stimulated Raman scattering and third-order dispersion effects induces only a small repetition rate mismatch Δf_{rep} of 20.6 MHz per 1.7 GHz of laser tuning, which is observed as acceleration and deceleration of the soliton in the cavity (see Fig. 2d, bottom). The fundamental limit for the tuning speed $d\Delta/dt < (\kappa/2\pi)^2$ is set by the cavity photon decay rate κ .

The linear dependence $df_{\text{rep}}/d\Delta \approx (\Omega_R/2\pi\Delta)(D_2/D_1)$, where $D_1/2\pi$ is the cavity free-spectral range and $D_2/2\pi$ is the second-order dispersion,

results in a channel-dependent frequency excursion B_μ and, hence, a constant rescaling factor of the measured lidar distance, which we can determine during calibration. Only nonlinear dependencies of the pulse repetition rate f_{rep} on the detuning Δ , from either the Raman shift³⁰ or multimodal interactions²⁹, actually degrade the linearity of the transduced chirp. The maximum detuning, which still supports stable DKS generation, is determined by the input pump power¹⁰, which in turn is fundamentally limited by a Raman instability³¹.

Characterization of parallel FMCW lidar source

Next, we experimentally demonstrate the ability to faithfully transfer the pump laser chirp to the individual comb teeth (see Fig. 3). Details of the experimental setup for heterodyne characterization, linearization of the triangular frequency-modulation patterns, and transduction data analysis are described in the Methods. Results for the comb tooth at 195 THz ($\mu = +20$) and modulation frequencies $1/T$ from 100 kHz to 10 MHz are depicted in Fig. 3b and in Extended Data Fig. 3. The frequency excursion bandwidth B_μ increases linearly with the channel number μ (see Fig. 3e) at a rate of $dB_\mu/d\mu = 22.15 \text{ MHz}$ in agreement with the predictions from numerical simulations including the Raman self-frequency

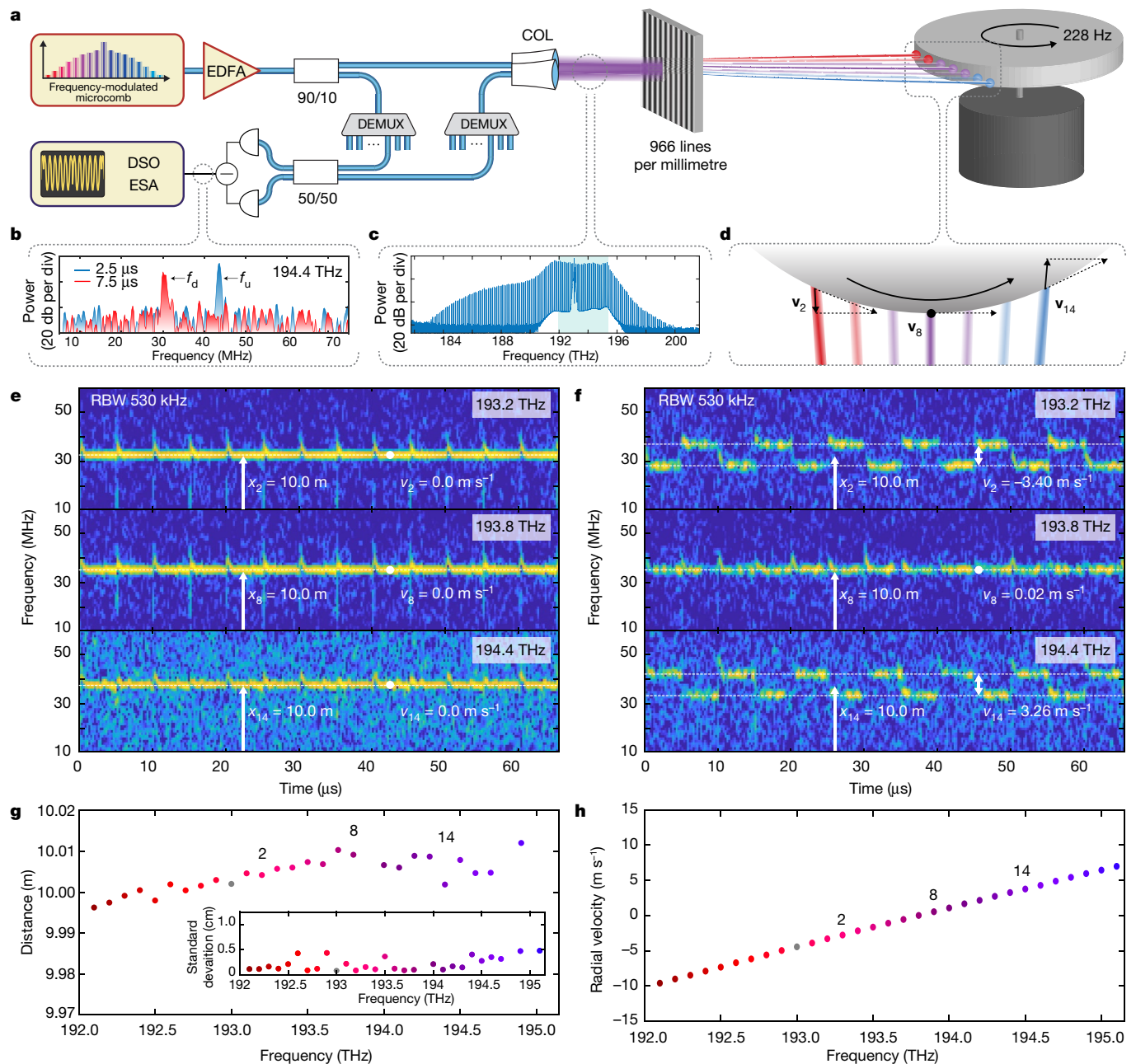


Fig. 4 | Demonstration of massively parallel velocity measurement using a soliton microcomb.

a, Experimental setup. The amplified frequency-modulated lidar microcomb source is split into signal and local oscillator pathways. The signal is dispersed with a transmission grating (966 lines per millimetre) over the horizontal circumference of a flywheel mounted on a small direct-current motor. The reflected signals are spectrally isolated before detection. COL, fibre collimator. **b**, Radio-frequency spectrum of lidar backreflection mixed with the local oscillator (sampling length 3.75 μs) around 2.5 μs (upward ramp) and 7.5 μs (downward ramp). **c**, Optical spectrum of comb lines after amplification. Blue shading highlights 30 comb lines with sufficient power (>0 dBm) for lidar detection. **d**, Schematic illustration of the flywheel

section irradiated by the frequency-modulated soliton microcomb lines indicating the projection of the position x_μ and velocity v_μ of the wheel onto the comb lines. **e**, Time–frequency maps of selected microcomb FMCW lidar channels (sampling length 0.5 μs) for the static flywheel. **f**, Same as **e**, but for flywheel rotating at 228 Hz. **g**, Multichannel distance measurement results for the static flywheel. Distance measurement not corrected for fibre path difference between signal and local oscillator path. **h**, Multichannel velocity measurement for the flywheel rotating at 228 Hz. The accuracy of distance and velocity measurements in case the rotating flywheel is affected by vibrations.

shift (see Fig. 2c). We define the chirp nonlinearity as the deviation of the measured instantaneous frequency from a perfectly symmetric triangular frequency-modulation scan, estimated with least-squares fitting, and depict results for the pump and two comb teeth in Fig. 3c (bottom). Narrow peaks of the chirp nonlinearity are attributed to single-mode dispersive waves²⁹. We do not observe intermode breathing of the soliton³² in the present system. The channel-dependent

root-mean-square nonlinearity is depicted in the inset of Fig. 3e and remains below 1/500 of the full frequency excursion for all channels at 100-kHz modulation frequency. The frequency-dependent transduction of the frequency modulation from the pump laser to the DKS teeth is calculated from the transduced chirps (see Extended Data Fig. 3) and is plotted in Fig. 3d. We find a lower bound for the 3-dB modulation frequency cutoff of 40 MHz, which corresponds to a maximum

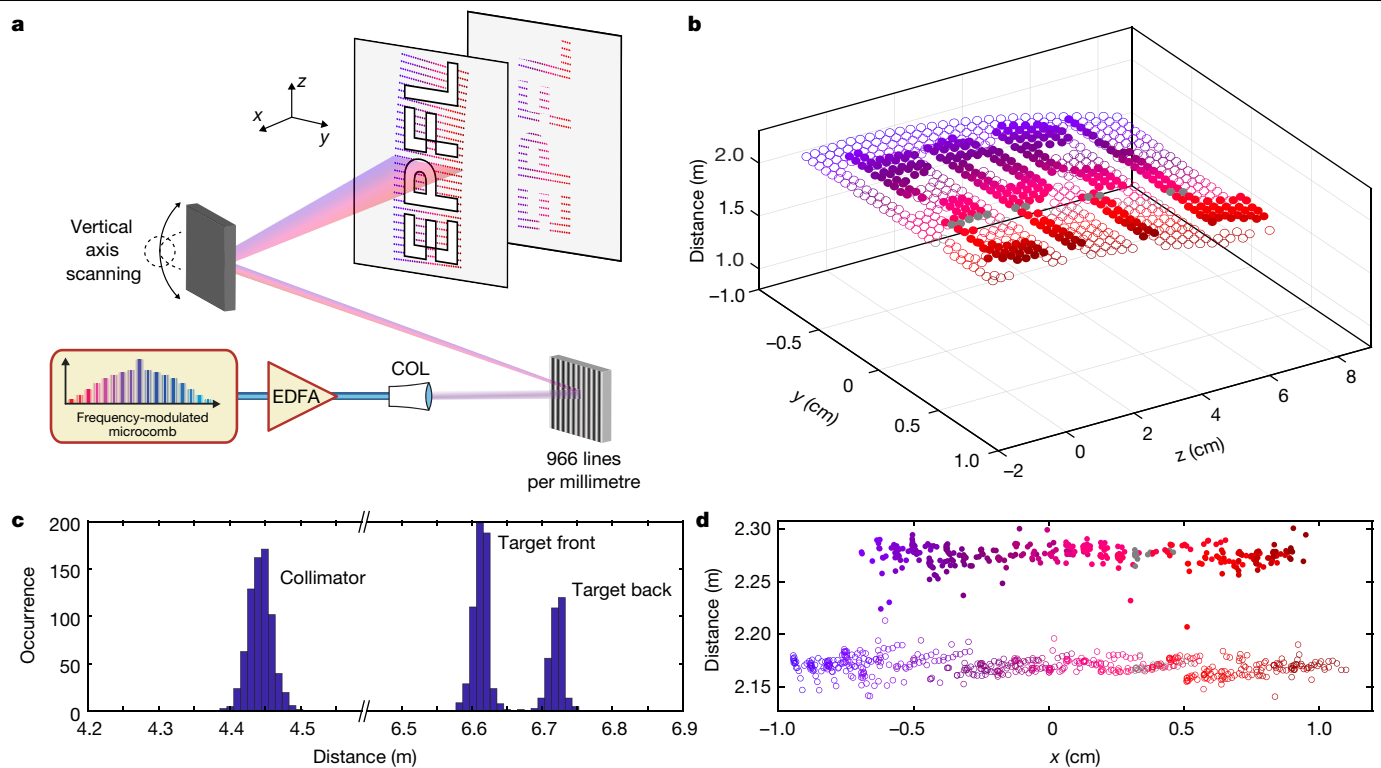


Fig. 5 | Parallel distance measurement and imaging. **a**, Experimental setup. 30 channels of the soliton microcomb are spectrally dispersed with a transmission grating in the horizontal axis (y). Vertical translation is performed by a planar mirror placed behind the grating. The target is formed by two vertical sheets of paper placed at a distance of 11.5 cm. The EPFL university logo is cut out from the first sheet. The coloured dots mark the approximate positions of the individual beams during the scan and denote the

individual spectral channels according to Fig. 3e. **b**, A 3D image obtained by scanning the beam array in the vertical direction. Filled circles denote pixels detected in the target back plane. **c**, Histogram of successful detections for the collimator (zero distance plane in **b**, **d**), target front and back planes. **d**, Projection of **b** along the z axis reveals the centimetre-level distance measurement accuracy and precision for the 30 frequency-modulated lidar channels.

per-channel chirp rate of $1.6 \times 10^{17} \text{ Hz}^2$. The estimated accumulated chirp rate of all channels thus rivals state-of-the-art swept source lasers, which achieve chirp rates of 10^{18} – 10^{19} Hz^2 (ref. ³³), recently used in time-stretch time-of-flight lidar³⁴.

Parallel ranging, velocimetry and 3D imaging

Next, we perform a proof-of-concept demonstration of the massively parallel lidar system. The calibrated frequency-modulated microcomb is split (90/10) into a signal path, which is spectrally dispersed around the circumference of the flywheel by a transmission grating (966 lines per millimetre), and a local oscillator path. The spectral channels of the reflected signal and the local oscillator are isolated using a bidirectional arrayed waveguide grating. The results of parallel distance and velocity measurement including standard deviations over 100 frequency-modulated periods for the static wheel are displayed in Fig. 4e, g. Channels beyond 195.2 THz are not observed with sufficient signal-to-noise ratio, because of limited amplification bandwidth. The measurement imprecision over 25 spectral channels is below 1 cm, comparable with state-of-the-art time-of-flight lidar systems and can be improved by using more broadband chirps. Small systematic offsets on the centimetre scale are associated with the lengths of the fibre pigtailed in the demultiplexers and switches. The results for the wheel spinning at 228 Hz are depicted in Fig. 4f, resolving the position dependency of the projected velocity around the circumference of the wheel (see Fig. 4d). The measurement accuracy in case of the spinning wheel is limited by vibration. The equivalent distance and velocity sampling rate of the 30 independent channels is 3 megapixels per second.

Last, we demonstrate parallel 3D imaging of 30 channels spectrally dispersed with a transmission grating and concurrently illuminating a target composed of two sheets of white paper spaced by 11 cm with the EPFL university logo cutout in the front plane (see Fig. 5). The target profile is imaged by translation of the beams in the vertical direction with a 45° steering mirror, and depicted in Fig. 5b. The detection is monostatic and the co-observed backreflection from the collimation lens serves as the zero-distance plane in the measurement. Target points detected in the back plane are clearly separated, owing to the centimetre-level distance precision and accuracy observed on all 30 FMCW channels (see Fig. 5c, d) and highlighted as filled points.

Discussion and conclusion

Thus we have described a method for massively parallel coherent lidar using photonic chip-based soliton microcombs. It enables us to reproduce arbitrary frequency chirps of the narrow linewidth pump laser onto all comb teeth that compose the soliton at speeds beyond 10^{17} Hz^2 , and has the potential to greatly increase the frame rate of imaging coherent lidar systems via parallelization. In contrast to earlier works in frequency-comb-based lidar^{21–23}, the comb teeth in parallel FMCW lidar are spatially dispersed with diffractive optics and separately measure distances and velocities in a truly parallel fashion. Assuming a setup similar to that of ref. ³⁵, that is, 179 carriers with 50-GHz spacing in the C+L telecommunications wavelength bands (1,530–1,610 nm), we expect aggregate pixel measurement rates of 17.9 megapixels per second for 100-kHz modulation frequency and 179 megapixels per second for 1-MHz modulation frequency, well beyond the present technologies of long-range time-of-flight and FMCW lidar systems.

Although the residual nonlinearity and slow power modulation of the comb sidebands during the frequency chirp only weakly influences the distance and velocity evaluation, we emphasize that both effects can be avoided entirely if both the laser and cavity are modulated in unison³⁶. Similarly, the laser can be self-injection-locked to the modulated cavity, which can furthermore extend the laser coherence length substantially^{37,38}. Promising actuation technologies include recently developed high-bandwidth and energy-efficient integrated electro-optical³⁹ and piezoelectrical actuators³⁶.

Moreover, by virtue of the laser line separations, our concept is compatible with nanophotonics-based gratings for beam separation and could greatly simplify optical phased array systems¹², wherein one axis of beam separation is provided by the nanophotonic grating and a second axis is provided by integrated phase shifters. Furthermore, this concept alleviates problems with eye safety, as the light is dispersed over multiple detection pixels at all times, similar to time-of-flight flash systems, yet avoids the problems associated with the excessive peak powers of high-energy pulsed light sources. Finally, spectrally multiplexed detection can also be carried out in a dual-comb approach, whereby the second comb scans in unison with the first, but has a different repetition rate, which removes the need for demultiplexing and individual detection of the comb lines. It should be noted that (resonant) electro-optical frequency combs^{39,40} based on LiNbO₃ also provide a platform in which the approach presented here could be realized. Hence, we conclude that microcombs, combined with concurrent advances in chip-scale lasers, optical beamforming structures, and hybrid electro-optical integration provide a path towards rapid, precise and simultaneously long-range coherent lidar modules suitable for industrial, automotive and airborne applications demanding high-speed 3D imaging in excess of ten megapixels per second.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2239-3>.

1. Bostick, H. A carbon dioxide laser radar system. *IEEE J. Quantum Electron.* **3**, 232 (1967).
2. Urmson, C. et al. Autonomous driving in urban environments: Boss and the urban challenge. *J. Field Robot.* **25**, 425–466 (2008).
3. Behroozpour, B., Sandborn, P., Wu, M. & Boser, B. E. Lidar system architectures and circuits. *IEEE Commun. Mag.* **55**, 135–142 (2017).
4. MacDonald, R. I. Frequency domain optical reflectometer. *Appl. Opt.* **20**, 1840–1844 (1981).
5. Uttam, D. & Culshaw, B. Precision time domain reflectometry in optical fiber systems using a frequency modulated continuous wave ranging technique. *J. Lightwave Technol.* **3**, 971–977 (1985).
6. Gnanalingam, S. & Weekes, K. Weak echoes from the ionosphere with radio waves of frequency 1.42 Mc./s. *Nature* **170**, 113–114 (1952).
7. Hymans, A. J. & Lait, J. Analysis of a frequency-modulated continuous-wave ranging system. *Proc. IEE B* **107**, 365 (1960).
8. Roos, P. A. et al. Ultrabroadband optical chirp linearization for precision metrology applications. *Opt. Lett.* **34**, 3692–3694 (2009).
9. Kippenberg, T. J., Gaeta, A. L., Lipson, M. & Gorodetsky, M. L. Dissipative Kerr solitons in optical microresonators. *Science* **361**, ean8083 (2018).

10. Lucas, E., Guo, H., Jost, J., Karpov, M. & Kippenberg, T. J. Detuning-dependent properties and dispersion-induced instabilities of temporal dissipative Kerr solitons in optical microresonators. *Phys. Rev. A* **95**, 043822 (2017).
11. McManamon, P. F. et al. Optical phased array technology. *Proc. IEEE* **84**, 268–298 (1996).
12. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493**, 195–199 (2013).
13. Levinson, J. et al. Towards fully autonomous driving: systems and algorithms. *Proc. IEEE Intelligent Vehicles Symp.* 163–168, <https://doi.org/10.1109/IVS.2011.5940562> (2011).
14. Maddern, W., Pascoe, G., Linegar, C. & Newman, P. 1 year, 1000 km: the Oxford robotcar dataset. *Int. J. Robot. Res.* **36**, 3–15 (2017).
15. Bosch, T. Laser ranging: a critical review of usual techniques for distance measurement. *Opt. Eng.* **40**, 10 (2001).
16. Schwarz, B. Mapping the world in 3D. *Nat. Photonics* **4**, 429–430 (2010).
17. Mitchell, E. W. et al. Coherent laser ranging for precision imaging through flames. *Optica* **5**, 988 (2018).
18. Petit, J., Stottelaar, B., Feiri, M. & Kargl, F. Remote attacks on automated vehicles sensors: experiments on camera and LiDAR. *Black Hat Europe Conf.* **11**, 1–13 (2015); <https://www.blackhat.com/docs/eu-15/materials/eu-15-Petit-Self-Driving-And-Connected-Cars-Fooling-Sensors-And-Tracking-Drivers-wp1.pdf>.
19. Herr, T. et al. Temporal solitons in optical microresonators. *Nat. Photonics* **8**, 145–152 (2014).
20. Leo, F. et al. Temporal cavity solitons in one-dimensional Kerr media as bits in an all-optical buffer. *Nat. Photonics* **4**, 471–476 (2010).
21. Suh, M. & Vahala, K. J. Soliton microcomb range measurement. *Science* **359**, 884–887 (2018).
22. Trocha, P. et al. Ultrafast optical ranging using microresonator soliton frequency combs. *Science* **359**, 887–891 (2018).
23. Kuse, N. & Fermann, M. Frequency-modulated comb LiDAR. *APL Photonics* **4**, 106105 (2019).
24. Pfeiffer, M. H. P. et al. Photonic damascene process for integrated high-Q microresonator based nonlinear photonics. *Optica* **3**, 20–25 (2016).
25. Karpov, M. et al. Raman self-frequency shift of dissipative Kerr solitons in an optical microresonator. *Phys. Rev. Lett.* **116**, 103902 (2016).
26. Guo, H. et al. Universal dynamics and deterministic switching of dissipative Kerr solitons in optical microresonators. *Nat. Phys.* **13**, 94–102 (2017).
27. Lugiato, L. A. & Lefever, R. Spatial dissipative structures in passive optical systems. *Phys. Rev. Lett.* **58**, 2209–2211 (1987).
28. Chembo, Y. K. & Menyuk, C. R. Spatiotemporal Lugiato-Lefever formalism for Kerr-comb generation in whispering-gallery-mode resonators. *Phys. Rev. A* **87**, 053852 (2013).
29. Yi, X. et al. Single-mode dispersive waves and soliton microcomb dynamics. *Nat. Commun.* **8**, 14869 (2017).
30. Yi, X., Yang, Q.-F., Yang, K. Y. & Vahala, K. Theory and measurement of the soliton self-frequency shift and efficiency in optical microcavities: publisher's note. *Opt. Lett.* **41**, 3722 (2016).
31. Wang, Y., Anderson, M., Coen, S., Murdoch, S. G. & Erkintalo, M. Stimulated Raman scattering imposes fundamental limits to the duration and bandwidth of temporal cavity solitons. *Phys. Rev. Lett.* **120**, 053902 (2018).
32. Guo, H. et al. Intermodal breather solitons in optical microresonators. *Phys. Rev. X* **7**, 041055 (2017).
33. Klein, T. et al. Multi-MHz retinal OCT. *Biomed. Opt. Express* **4**, 1890 (2013).
34. Jiang, Y., Karpf, S. & Jalali, B. Time-stretch LiDAR as a spectrally scanned time-of-flight ranging camera. *Nat. Photonics* **14**, 14–18 (2020).
35. Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. *Nature* **546**, 274–279 (2017).
36. Liu, J. et al. Monolithic piezoelectric control of soliton microcombs. Preprint at <https://arxiv.org/abs/1912.08686> (2020).
37. Liang, W. et al. High spectral purity Kerr frequency comb radio frequency photonic oscillator. *Nat. Commun.* **6**, 7957 (2015).
38. Pavlov, N. et al. Narrow-linewidth lasing and soliton Kerr microcombs with ordinary laser diodes. *Nat. Photonics* **12**, 694–698 (2018).
39. Zhang, M. et al. Broadband electro-optic frequency comb generation in a lithium niobate microring resonator. *Nature* **568**, 373–377 (2019).
40. Metcalf, A. J., Torres-Company, V., Leaird, D. E. & Weiner, A. M. High-power broadly tunable electrooptic frequency comb generator. *IEEE J. Sel. Top. Quantum Electron.* **19**, 231–236 (2013).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Sample details and fabrication

Integrated Si_3N_4 microresonators are fabricated with the photonic damascene process⁴¹, deep-ultraviolet stepper lithography⁴² and silica preform reflow⁴³. The waveguide cross-section is 1.5 μm wide and 0.82 μm high, with anomalous second-order dispersion of $D_2/2\pi = 1.13$ MHz and third-order dispersion parameter of $D_3/2\pi = 576$ Hz, where the positions of the resonance frequencies close to the pumped resonance are expressed by the series $\omega_\mu = \omega_0 + \sum_{i \geq 1} D_i \mu_i / i!$. The ring radius is 228.43 μm and results in a resonator free-spectral-range of $D_1/2\pi = 98.9$ GHz, which is chosen to match the standard 100-GHz telecommunication channel grid. The resonator is operated in the strongly overcoupled regime with an intrinsic loss rate $\kappa_0/2\pi = 15$ MHz and bus waveguide coupling rate $\kappa_{\text{ex}}/2\pi = 100$ MHz. Operation in the strongly overcoupled regime has the advantage of suppressing thermal nonlinearities during tuning as well as increasing the power per comb line before and optical signal-to-noise ratio after post-amplification. Input and output coupling of light to and from the fundamental transverse electric (TE) mode of the photonic chip is facilitated with double inverse tapers⁴⁴ and lensed fibres.

Frequency-modulated soliton microcomb generation

We set up a frequency-agile pump laser for soliton generation using a continuous-wave external cavity diode laser coupled into an electro-optical phase modulator for measurement of the relative laser cavity detuning, and dual Mach–Zehnder modulator (single sideband modulator) biased to single sideband modulation, which is driven by a frequency-agile voltage controlled oscillator (VCO, 5–10 GHz) and an arbitrary function generator. The continuous-wave laser is amplified to 1.7 W and 1 mW is split off for chirp linearization in a separate imbalanced Mach–Zehnder fibre interferometer (MZI) for chirp linearization purposes^{45,46}. The DKS^{19,20} is generated by coupling the frequency-modulated pump laser onto the photonic chip and tuning of the laser into resonance and single soliton state using the established piezo tuning scheme^{19,26}. The detuning with respect to the Kerr shifted cavity resonance and the bistable soliton response is monitored using a vector network analyser driving a weak phase modulation via an inline electro-optical-modulator²⁶ and an optical spectrum analyser. The generated soliton is coupled back into the optical fibre, the residual pump light is filtered and the soliton pulse train is amplified with a gain-flattening erbium-doped fibre amplifier. The repetition rate of the soliton pulse train is 99 GHz and the cavity resonance is aligned to the telecom channel C30 at a wavelength of 1,553.3 nm using a thermo-electric cooling device located below the active chip. Although it is possible to directly modulate all comb teeth post-DKS generation, this method suffers from excess insertion loss of the single sideband modulation around 15 dB, which severely degrades the optical signal-to-noise ratio after post-amplification. Even more critically, the internal MZIs of the single-sideband modulator are wavelength-dependent and require precise direct-current biasing to suppress the unwanted carrier and higher-order sidebands, which limits the usable optical bandwidth for direct single sideband modulation of the soliton comb. Our scheme of single sideband modulation of the soliton pump laser has the advantage that the fundamental carrier and unwanted sidebands are not transduced in the cavity and no sidebands on the comb teeth appear. Moreover, our scheme works irrespective of the choice of laser and microresonator actuation schemes, especially established FMCW sources that are based on frequency-agile diode lasers. The residual effects of the cavity, limiting both the chirp range and inducing a small nonlinearity in the transduction can be alleviated strongly by concurrent actuation of the cavity and the diode laser³⁹, with direct injection locking of the laser to the modulated cavity constituting an especially compact and technologically promising implementation scheme^{36,47}.

Linearization and calibration

Frequency-modulated lidar requires perfectly linear chirp ramps to achieve precise and accurate distance measurements⁸. We implemented a digital pre-distortion circuit in order to minimize the chirp nonlinearity of the pump frequency sweep, similar to prior implementations⁴⁸. The optimization procedure was applied in two configurations to measure the pump frequency chirp, either via heterodyne with a reference laser (see Fig. 3), or via delayed homodyne detection in an imbalanced MZI. The length difference of the calibration MZI arms (12.246 m) is determined using the electro-optical phase modulator and vector network analyser and fitting the \sin^2 spectral response function of the MZI. The setup and optimization results for this method are detailed in Extended Data Fig. 1. The chirp is applied to the continuous-wave laser with a VCO-driven single sideband modulator. The VCO is initially driven by a simple triangular function generated using the arbitrary function generator. The driving voltage is then iteratively corrected to improve the chirp linearity. After modulation, a fraction of the light is picked up to generate a beat note with a reference external cavity diode laser. The downmixed laser frequency is sampled on a digital sampling oscilloscope (20 gigasamples per second) and digitally processed to perform a short-term Fourier transform followed by peak detection. The measured frequency evolution is fitted with a perfect triangular function having a fixed target frequency excursion. This allows the deviation from this desired frequency chirp to be assessed. The frequency deviation is then converted to voltage—after computing the average voltage-to-frequency coefficient of the VCO—and then added to the current tuning function of the arbitrary function generator. This procedure effectively addresses the nonlinear response of the VCO, as shown in Extended Data Fig. 1b–e. The optimization procedure was applied successfully at different tuning speeds (10 kHz–10 MHz), as shown in Extended Data Fig. 2. However, with increasing tuning speed, the residual root-mean-square deviation increases, which we attribute to the limited tuning bandwidth of the VCO.

Heterodyne characterization of frequency-modulated soliton microcomb

Heterodyne characterization of the transduced modulation is carried out to avoid possible ambiguities of delayed homodyne detection and catch high-frequency noise components obscured in low bandwidth detection. The spectral channels are isolated using a commercial telecommunications wavelength-division demultiplexer based on planar arrayed waveguide gratings and superimposed on a high-bandwidth (10 GHz) balanced photoreceiver. The data are recorded on a high-bandwidth balanced photodetector and a fast realtime sampling oscilloscope. Modulation frequencies span from 10 kHz to 10 MHz in our study and are limited by the actuation bandwidth of the arbitrary function generator and VCO. The total measurement duration is between 0.5 ms (10 kHz) and 30 μs (10 MHz). The instantaneous frequency is determined via short-time Fourier transform using a 4th-order Nuttall window and in the case of the pump channel (193 THz) is linearized by applying iterative predistortion of the VCO input (see Extended Data Fig. 1). The resolution bandwidth Δf of the transform window is adjusted to minimize the effective linewidth of the chirped signal:

$$\Delta f = \sqrt{\frac{2B}{T}} \quad (1)$$

By tuning the second external cavity diode laser close to the individual comb teeth, we can separately measure the transduced frequency modulation patterns for each comb sideband within the bandwidth of the demultiplexer. The resulting time frequency maps for the modulation frequencies 100 kHz and 10 MHz across five modulation periods are depicted in Extended Data Fig. 7.

Article

The tuning nonlinearity of the comb teeth is calculated as the root-mean-square deviation of the measured tuning curve from a perfect triangular frequency modulation trace determined with least-squares fitting. We determine frequency-dependent transduction from the intensities of the 1st to 9th harmonic of the triangular frequency-modulated spectrum, which we normalize with respect to the corresponding pump modulation amplitude (see Extended Data Fig. 4). We observe a slight amplification of the modulation for frequencies around 100 MHz on both the pump and sideband. A fine analysis reveals three effects. For low modulation frequencies, weak even-order sidebands arise, which we attribute to the hysteresis effect, which accompanies the generation of single mode dispersive waves^{29,32}, essentially introducing a small asymmetry in the transduced chirp.

Parallel velocimetry and ranging

The experimental setup is illustrated in Fig. 4a. The frequency modulation $1/T$ and excursion B of the microcomb pump are adjusted to 100 kHz and 1.7 GHz, respectively. The frequency-modulated comb is amplified with a gain-flattened erbium-doped fibre amplifier and split into signal (90%) and local oscillator (10%) paths. A total power of 350 mW is emitted from the collimator, which equates to between 5 mW and 20 mW per comb line. A transmission grating (966 lines per millimetre) spectrally disperses the individual signal comb lines along the circumference of the flywheel. Normal incidence reflection of the wheel is obtained by the frequency-modulated microcomb sideband at 193.8 THz. A bistatic detection with separate collimators for the transmit and receive paths is chosen to minimize spurious back-reflection in the fibre components. The back-reflected signal and local oscillator comb lines are spectrally separated in the demultiplexer and superimposed on a balanced photodetector for detection. Two 1×40 mechanical optical switches are installed with the demultiplexers to allow individual channels to be measured sequentially, removing the requirement to provide 30 balanced photodetectors and analogue-to-digital converters. The total optical loss budget of the demultiplexing network is around 5 dB per channel and could be reduced by co-integration on the photonic chip. We stress that in all measurements illuminating and receiving light and demultiplexing the pixels are done simultaneously. Hence, any additional noise and crosstalk between the channels would be detected in our setup. However, our system is impervious to crosstalk and interference between the channels, because of the spectral channel separation, in contrast to simple spatial channel separation⁴⁹, which requires sequential operation. Our current setup utilizes discrete telecommunications fibre components and optical switches for the detection, but we emphasize that high-performance integrated photonic solutions for many-channel dense wavelength division multiplexing communications have been demonstrated⁵⁰ and can be integrated on the Si_3N_4 photonic chip with performance comparable to that of the commercial telecommunication components employed here^{51,52}. The calibration of the channel-dependent frequency excursion bandwidth for the ranging experiments is performed using a second MZI (8.075 m; see Extended Data Fig. 6). The calibration curve is detected once before the start of the measurements and assumed constant throughout. The distance and velocity precision and accuracy of the system are determined using a small flywheel (radius 20 mm) mounted on a fast direct-current motor spinning at up to 228 Hz (see Fig. 4). The data analysis is performed with a simple Fourier transform accounting for a constant 535 ns delay between the arbitrary function generator and the lidar lasers, which is mainly obtained from the optical fibre lengths of the erbium-doped fibre amplifiers. We apply a 4th-order Blackmann–Harris type window function with effective resolution bandwidth 530 kHz. This resolution bandwidth corresponds to a range bin width of 7.9 cm (192.1 THz) to 5.9 cm (194.9 THz), in accordance with the FMCW fundamental range resolution $\Delta x = c/2B$. Two spectra

corresponding to the upwards and downwards slopes of the frequency chirp are separately transformed within each FMCW period and we apply Gaussian peak fitting to determine the peak frequency in each spectrum. We determine the statistical distance error by calculating the standard deviation of 100 consecutive distance measurements, which deviate by less than 1 cm from the mean. Further improvements, especially in long-range detection, could be achieved using active demodulation analysis⁴⁶. The residual nonlinearity broadens the detected beat notes and reduces the signal-to-noise ratio and power efficiency of the system. We estimate that an optimized version of our architecture with concurrent tuning of the resonator and laser would feature improved precision and detection performance by up to one order of magnitude.

Demonstration of parallel imaging

The optical setup is depicted in Fig. 5a, wherein the optical receiver, demultiplexers and detectors are omitted for brevity, but are set up as depicted in Fig. 4a. The target is composed of two sheets of white paper spaced by 11.5 cm. The EPFL university logo (width 7.5 cm, height 22.7 cm) is cut from the first sheet and oriented vertically. The FMCW lidar channels are dispersed horizontally using a transmission grating of 966 lines per millimetre and directed to the target with a 45° steering mirror. A monostatic detection scheme using an optical circulator and single collimator (Fig. 1) is chosen. The detector aperture is increased by placing a 75-cm focal length lens 1 m away from the 4 mm collimator and behind the grating. We note that modal interactions with the fundamental transverse magnetic (TM) mode strongly increase the power fluctuation of channels at 195.2 THz and 195.3 THz and spoils their use in frequency-modulated lidar experiments by shortening the effective sampling length.

Data availability

The data used to produce the plots within this paper are available at <https://doi.org/10.5281/zenodo.3603614>.

Code availability

The code used to produce the plots within this paper is available at <https://doi.org/10.5281/zenodo.3603614>.

41. Pfeiffer, M. H. P. et al. Photonic damascene process for low-loss, high-confinement silicon nitride waveguides. *IEEE J. Sel. Top. Quantum Electron.* **24**, 1–11 (2018).
42. Liu, J. et al. Ultralow-power chip-based soliton microcombs for photonic integration. *Optica* **5**, 1347–1353 (2018).
43. Pfeiffer, M. H. P. et al. Ultra-smooth silicon nitride waveguides based on the damascene reflow process: fabrication and loss origins. *Optica* **5**, 884–892 (2018).
44. Liu, J. et al. Double inverse nanotapers for efficient light coupling to integrated photonic devices. *Opt. Lett.* **43**, 3200–3203 (2018).
45. Ahn, T. J. & Kim, D. Y. Analysis of nonlinear frequency sweep in high-speed tunable laser sources using a self-homodyne measurement and hilbert transformation. *Appl. Opt.* **46**, 2394 (2007).
46. Feneyrou, P. et al. Frequency-modulated multifunction LiDAR for anemometry, range finding, and velocimetry: 1. Theory and signal processing. *Appl. Opt.* **56**, 9663 (2017).
47. Shen, B. et al. Integrated turnkey soliton microcombs operated at CMOS frequencies. Preprint at <https://arxiv.org/abs/1911.02636> (2019).
48. Zhang, X., Pouls, J. & Wu, M. C. Laser frequency sweep linearization by iterative learning pre-distortion for FMCW lidar. *Opt. Express* **27**, 9965 (2019).
49. Martin, A. et al. Photonic integrated circuit-based FMCW coherent LiDAR. *J. Lightwave Technol.* **36**, 4640–4645 (2018).
50. Fang, Q. et al. WDM multi-channel silicon photonic receiver with 320 Gbps data transmission capability. *Opt. Express* **18**, 5106–5113 (2010).
51. Ahn, D. et al. High performance, waveguide integrated Ge photodetectors. *Opt. Express* **15**, 3916 (2007).
52. Piels, M., Bauters, J. F., Davenport, M. L., Heck, M. J. R. & Bowers, J. E. Low-loss silicon nitride AWG demultiplexer heterogeneously integrated with hybrid III-V/silicon photodetectors. *J. Lightwave Technol.* **32**, 817–823 (2014).

Acknowledgements We thank A. S. Raja for his contribution with microresonator testing. Samples were fabricated at the Center of MicroNanoTechnology (CMi) with the assistance of R. N. Wang. This work was supported by funding from the Swiss National Science Foundation under grant agreement number 165933 and by the Air Force Office of Scientific Research

(AFOSR), Air Force Material Command, USAF, under award number FA9550-15-1-0250. Sample fabrication and process development was funded by contract HR0011-15-C-055 (DODOS) from the Defense Advanced Research Projects Agency (DARPA), Microsystems Technology Office (MTO). J.R. and W.W. acknowledge support from the EUs H2020 research and innovation program under the Marie Skłodowska-Curie IF grant agreement numbers 846737 (CoSiLiS) and 753749 (SOLISYNTH), respectively. We acknowledge interactions with A. Zott from ZEISS AG.

Author contributions A.L. and J.R. conducted the various experiments and analysed the data. E.L. assisted with laser linearization, W.W. performed the numerical simulations, A.L. designed

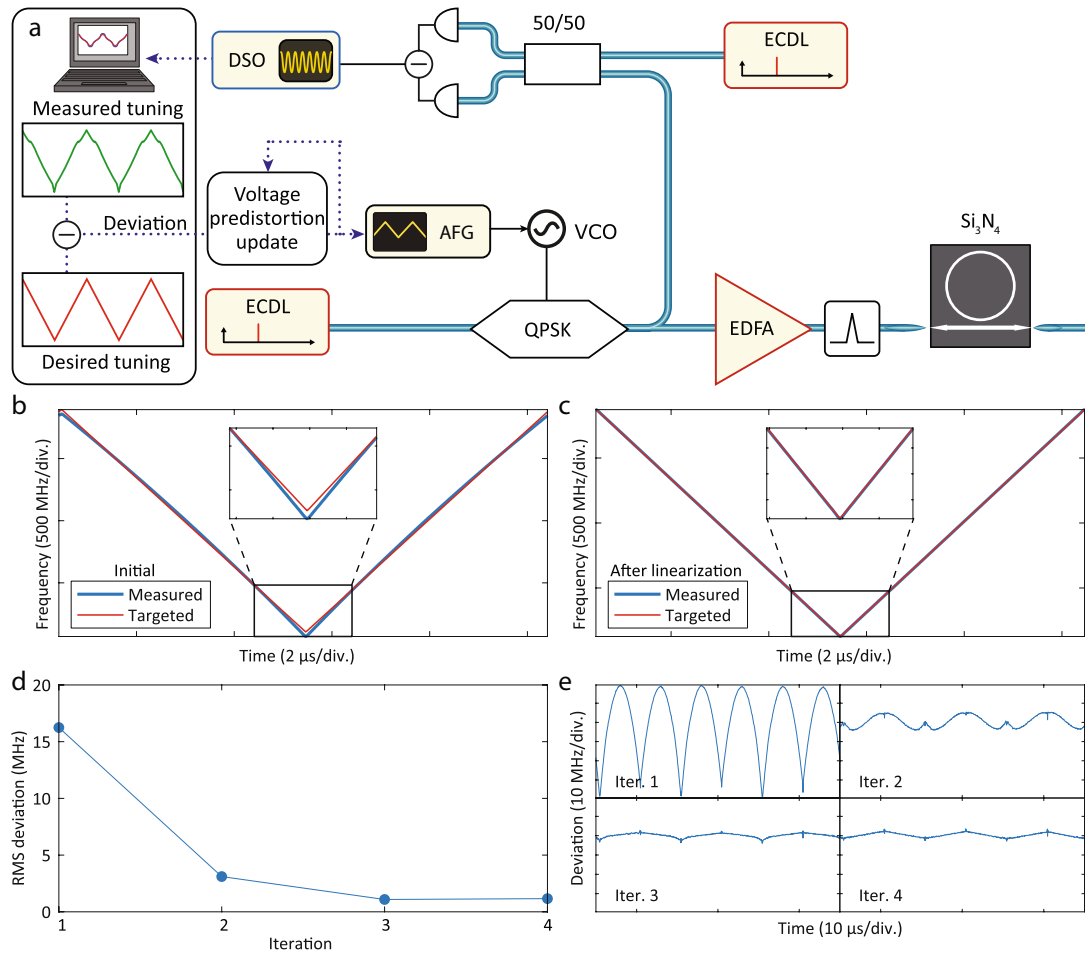
the samples and J.L. fabricated the samples. All authors discussed the manuscript. J.R., T.J.K., M.K. and E.L. wrote the manuscript. T.J.K. supervised the work and conceived the experiment.

Competing interests T.J.K. is a co-founder and shareholder of LiGenTec SA, a start-up company that is engaged in making Si_3N_4 nonlinear photonic chips available via foundry service.

Additional information

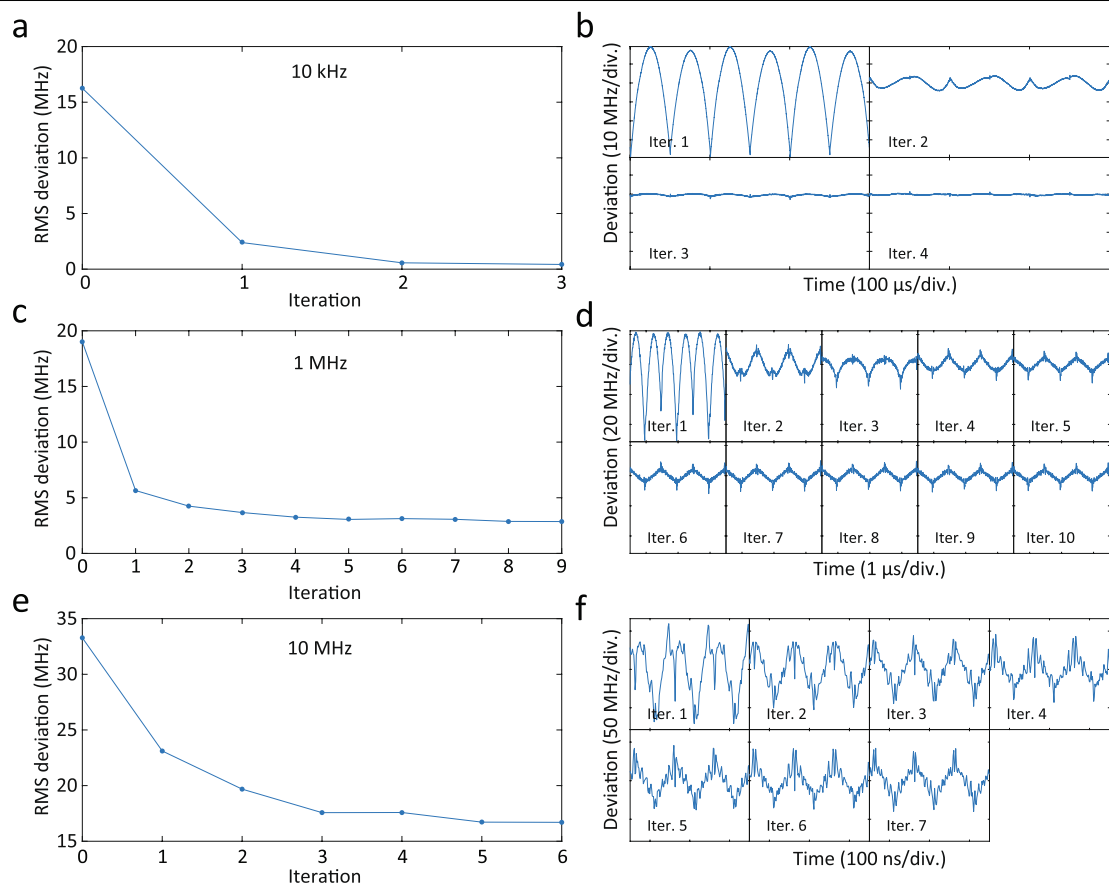
Correspondence and requests for materials should be addressed to T.J.K.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



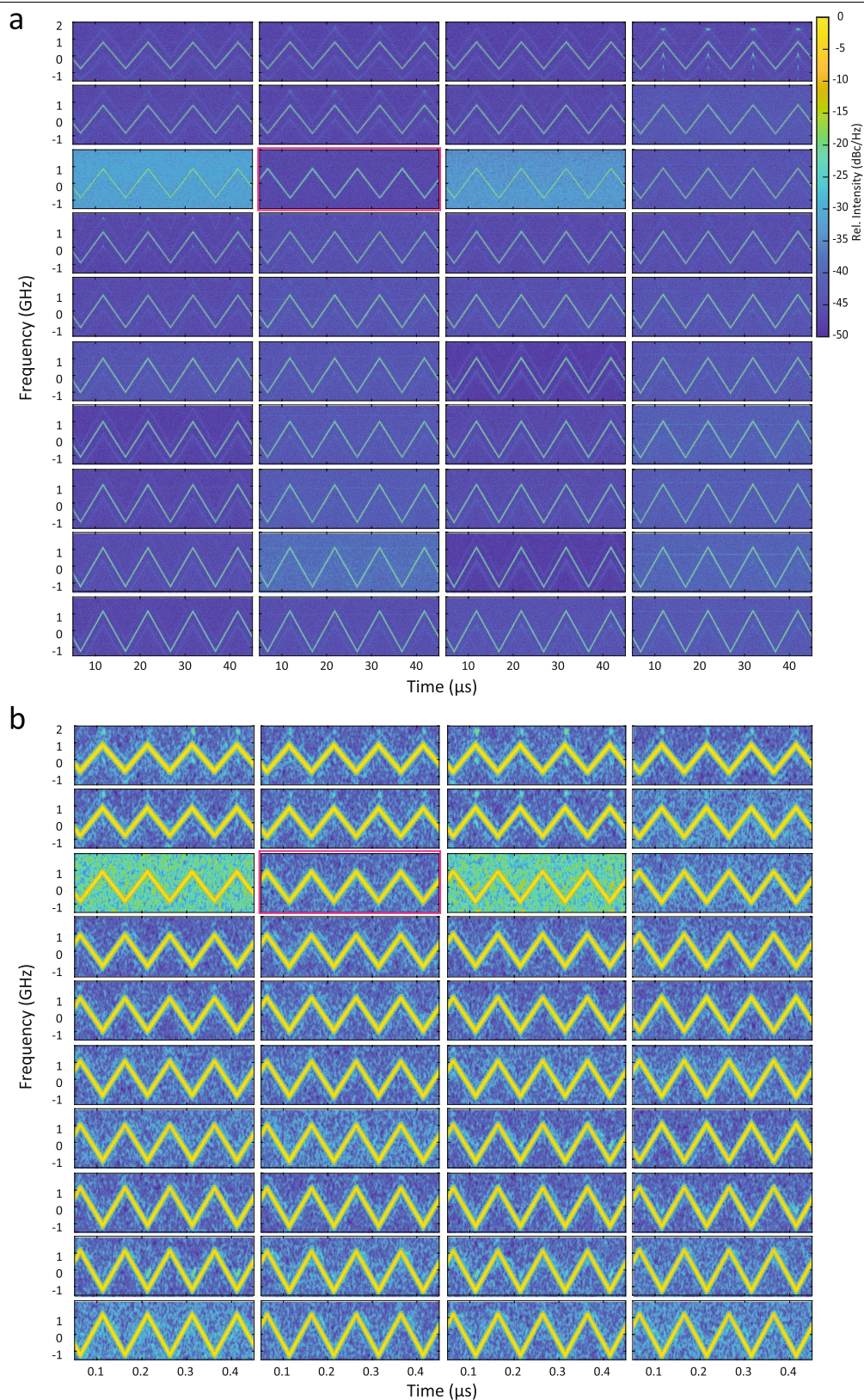
Extended Data Fig. 1 | Pump frequency sweep linearization via the heterodyne method. a, Setup for pump-laser frequency measurement via heterodyne beat note and chirp linearization feedback. **b**, Initial frequency modulation, when the VCO is driven with a triangular ramp. The measured frequency is compared with the targeted ideal modulation. The ramp

frequency is 100 kHz. **c**, Final triangular frequency modulation pattern, after four iterations. **d**, Evolution of the root-mean-square (RMS) frequency deviation during the optimization loop. **e**, Evolution of the deviation between measurement and target sweep, at each iteration of the loop.



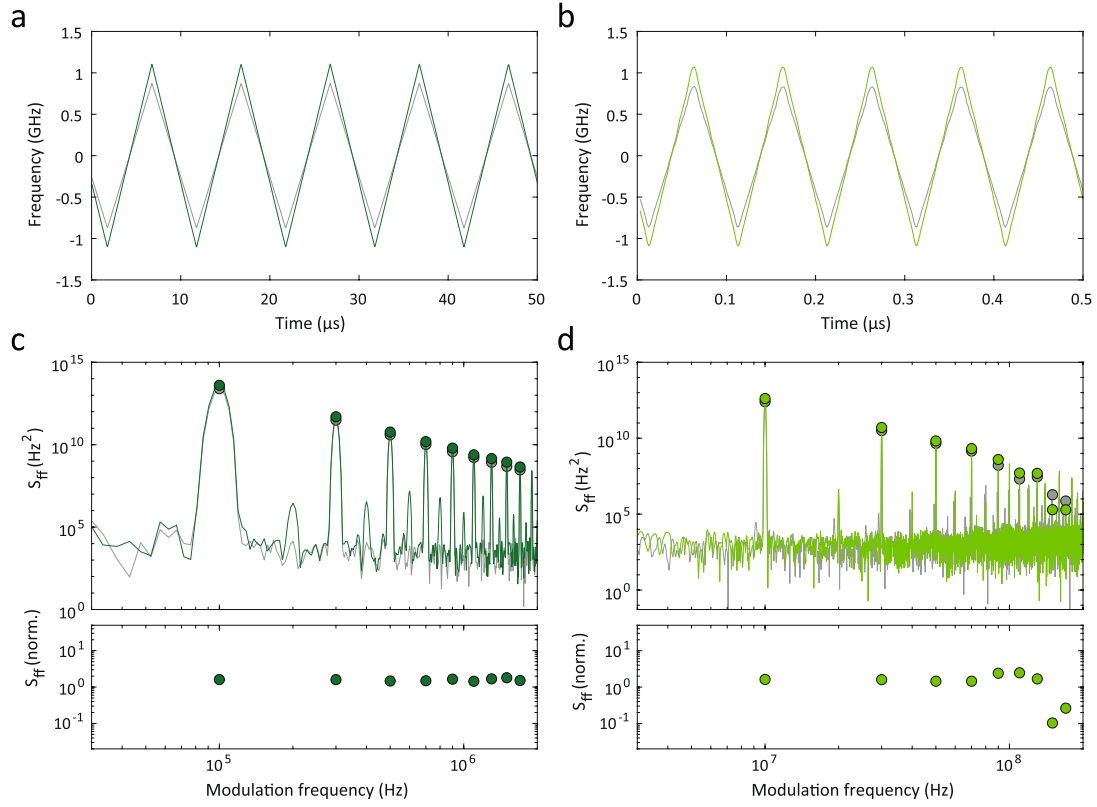
Extended Data Fig. 2 | Linearization results at different modulation frequencies. a, c, e. The evolution of the root-mean-square frequency deviation during the optimization loop for modulation frequencies of 10 kHz,

1 MHz and 10 MHz, respectively. **b, d, f.** Corresponding evolution of the deviation between the measurement and the target sweep, at each iteration of the loop.



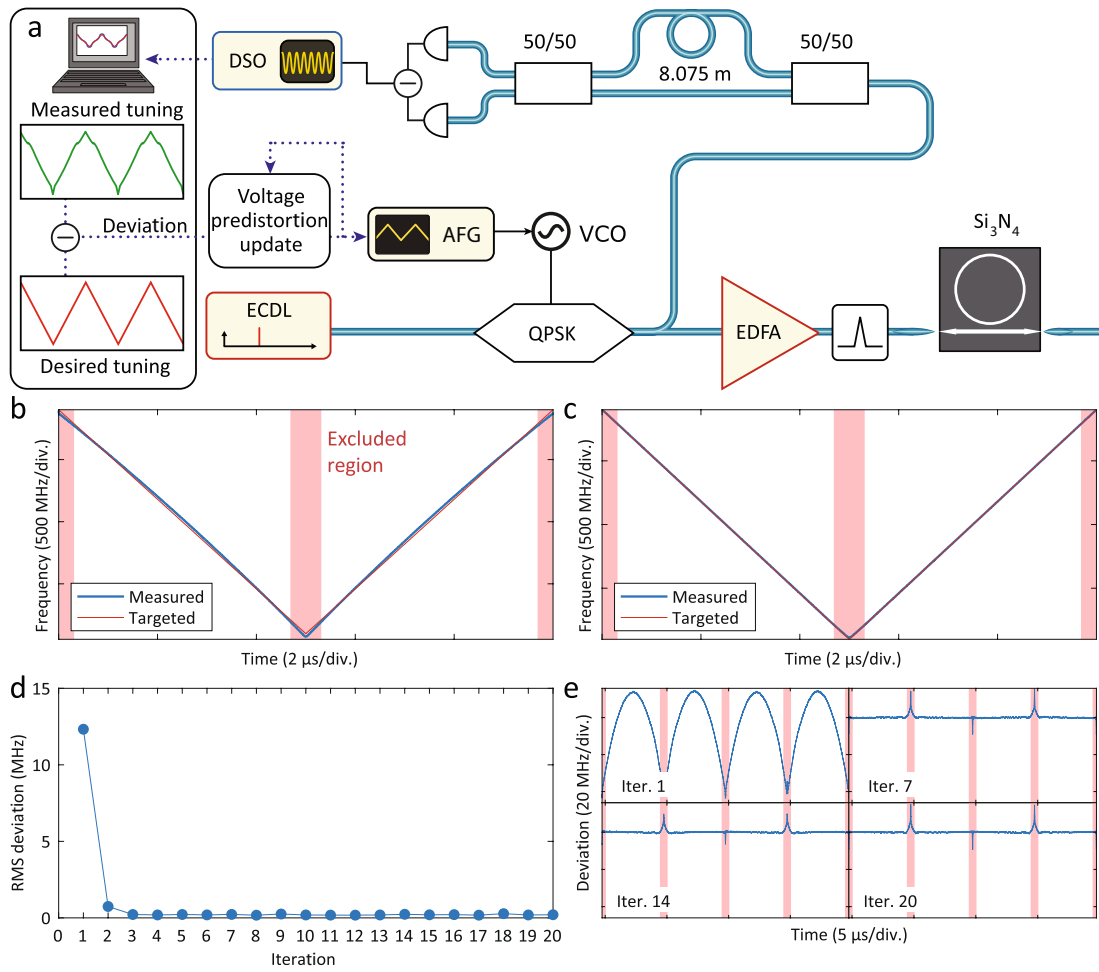
Extended Data Fig. 3 | Channel-by-channel analysis of heterodyne chirp characterization. a, Time–frequency maps obtained with short-time Fourier transform of the heterodyne beat detection of the individual FMCW channels.

Top left to bottom right panels denote optical carriers between 192.1 THz and 196 THz. Modulation frequency is 100 kHz. The pump channel at 193 THz is outlined in purple. **b,** As for **a**, but for modulation frequency 10 MHz.



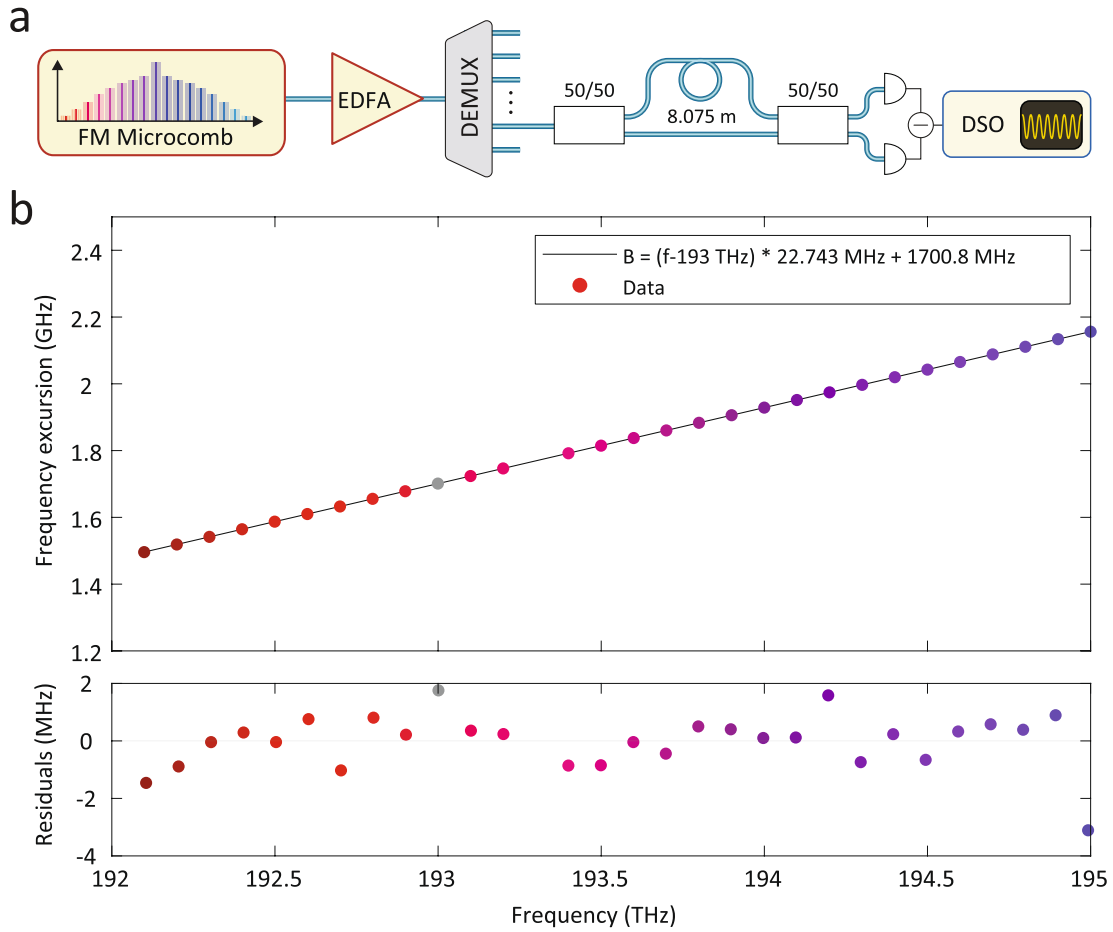
Extended Data Fig. 4 | Frequency-dependent transduction of carrier modulation from pump to comb sidebands. a, Time-dependent frequency of pump laser at 193 THz (grey) and 195 THz comb sideband ($\mu = 20$, dark green) and modulation frequency 100 kHz. **b**, As for **a**, but for modulation frequency 10 MHz. **c**, Power spectral density of frequency modulation S_{ff} for pump (grey)

and sideband (dark green). The markers denote the positions of harmonics, which are used in the transduction analysis. The lower panel shows the power spectral density of sideband frequency modulation harmonics normalized to the corresponding modulation power spectral density of the pump laser (see Fig. 3). **d**, As for **c**, but for modulation frequency 10 MHz.



Extended Data Fig. 5 | Pump frequency sweep linearization via the delayed homodyne method. **a**, Setup for pump-laser frequency measurement via delayed homodyne detection and chirp linearization feedback. Calibration of the MZI is performed by fitting the frequency-dependent phase modulation response of the MZI. **b**, Initial frequency modulation, when the VCO is driven with a triangular ramp, determined using a Hilbert transform. The measured frequency is compared with the targeted ideal modulation. The ramp

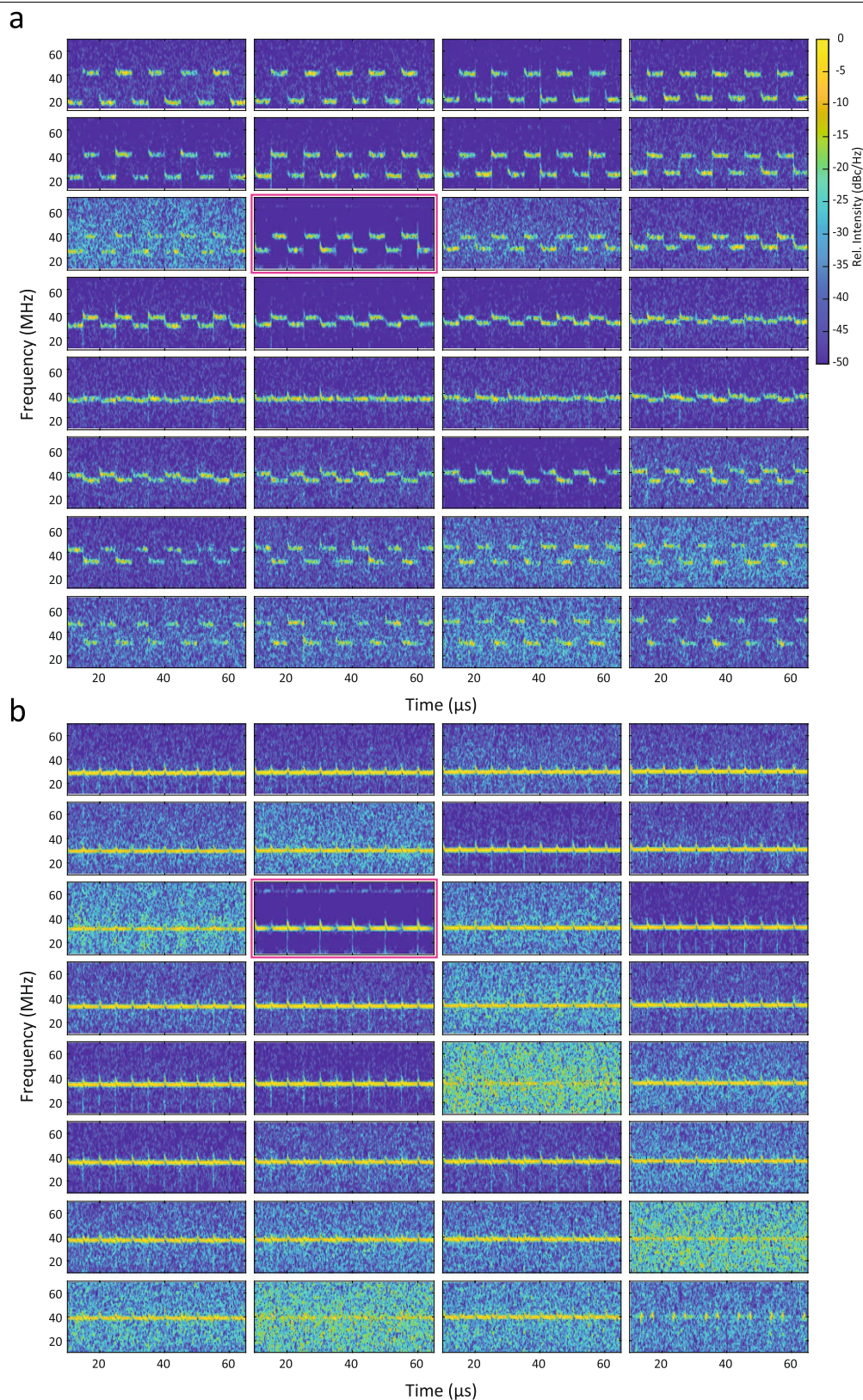
frequency is 100 kHz. The red-shaded regions around the extremal points are excluded from the linearization update. **c**, Final triangular frequency modulation pattern, after 20 iterations. Convergence achieved after four iterations. **d**, Evolution of the root-mean-square frequency deviation during the optimization loop. **e**, Evolution of the deviation between measurement and target sweep, at each iteration of the loop.



Extended Data Fig. 6 | Calibration of channel dependent frequency excursion bandwidth for distance and velocity measurements.

a, Measurement setup. The linearized frequency-modulated microcomb (see Extended Data Fig. 5 for setup schematic) is amplified and individual channels are selected by connecting the local oscillator path of the measurement setup

to a calibrated imbalanced MZI (8.075 m). **b**, The top panel shows the frequency-excursion bandwidth B_μ determined from independent measurement of the length of imbalanced MZI. Linear fit related to Raman self-frequency shift Ω_R . The bottom panel shows the residuals of the linear fit.



Extended Data Fig. 7 | Channel-by-channel analysis of proof-of-concept lidar demonstration. a, Time-frequency maps obtained with short-time Fourier transform of the delayed homodyne beat detection of the individual FMCW channels back-reflected from the rotating flywheel. Top left to bottom

right panels denote optical carriers between 192.1 THz and 195.2 THz. The pump channel at 193 THz is outlined in purple. Modulation frequency is 100 kHz. **b,** As for **a**, but for static flywheel.

Engineering covalently bonded 2D layered materials by self-intercalation

<https://doi.org/10.1038/s41586-020-2241-9>

Received: 2 November 2019

Accepted: 4 March 2020

Published online: 13 May 2020

 Check for updates

Xiaoxu Zhao^{1,2,9}, Peng Song^{2,9}, Chengcai Wang³, Anders C. Riis-Jensen⁴, Wei Fu², Ya Deng⁵, Dongyang Wan⁶, Lixing Kang⁵, Shoucong Ning¹, Jiadong Dan¹, T. Venkatesan^{1,6}, Zheng Liu⁵, Wu Zhou⁷, Kristian S. Thygesen⁴, Xin Luo⁸✉, Stephen J. Pennycook¹✉ & Kian Ping Loh²✉

Two-dimensional (2D) materials^{1–5} offer a unique platform from which to explore the physics of topology and many-body phenomena. New properties can be generated by filling the van der Waals gap of 2D materials with intercalants^{6,7}; however, post-growth intercalation has usually been limited to alkali metals^{8–10}. Here we show that the self-intercalation of native atoms^{11,12} into bilayer transition metal dichalcogenides during growth generates a class of ultrathin, covalently bonded materials, which we name ic-2D. The stoichiometry of these materials is defined by periodic occupancy patterns of the octahedral vacancy sites in the van der Waals gap, and their properties can be tuned by varying the coverage and the spatial arrangement of the filled sites^{7,13}. By performing growth under high metal chemical potential^{14,15} we can access a range of tantalum-intercalated TaS(Se)_y, including 25% Ta-intercalated Ta₉S₁₆, 33.3% Ta-intercalated Ta₇S₁₂, 50% Ta-intercalated Ta₁₀S₁₆, 66.7% Ta-intercalated Ta₈Se₁₂ (which forms a Kagome lattice) and 100% Ta-intercalated Ta₉Se₁₂. Ferromagnetic order was detected in some of these intercalated phases. We also demonstrate that self-intercalated V₁₁S₁₆, In₁₁Se₁₆ and Fe_xTe_y can be grown under metal-rich conditions. Our work establishes self-intercalation as an approach through which to grow a new class of 2D materials with stoichiometry- or composition-dependent properties.

Increased research into 2D materials has heralded a new branch of condensed-matter physics concerned with the description of electrons in atomically thin structures. So far, research efforts have primarily focused on 2D monolayers² and their hetero-stacked structures³, in which new properties can be engineered by generating superlattices of different moiré wavelengths. However, these hetero-stacked structures are currently produced by bottom-up methods that are low yielding and show poor reproducibility¹⁶. An alternative method of compositional tuning involves the intercalation of foreign atoms into the van der Waals (vdW) gap that is sandwiched by the chalcogen atoms; this has been shown to induce pseudo-2D characteristics in bulk crystals and modify their electronic properties^{4,6,7}. Depending on the interlayer stacking registries, the vdW gaps in transition metal dichalcogenides (TMDs) contain either octahedral and tetrahedral vacancies or trigonal-prismatic vacancies¹³, which provide docking sites for a diverse range of intercalants. Examples of successful intercalants include alkali metals^{8–10} such as Li, Na and K; transition metals^{17–21} such as Cu, Co, Ni, Fe and Nb; noble metals^{22–24} such as Ag, Au and Pt; as well as Sn and various organic molecules^{25–27}. Charge transfer from the intercalants⁷—or increased spin–orbit coupling due to the presence of heavy atoms^{7,24,28}—can enhance superconductivity¹⁰,

thermoelectricity²⁵ or spin polarization⁷. Intercalation is typically achieved using post-growth, diffusion-limited processes, either electrochemical or in the solid state. A well-defined intercalated phase with long-range crystalline order is difficult to obtain by such methods and usually requires harsh treatment conditions^{21,22,29}. Moreover, an intercalation phase diagram that correlates the density and spatial distribution of the intercalated atoms with the mesoscopic properties of the intercalation compound is currently lacking. Compared with the intercalation of foreign atoms into a TMD, the intercalation of native atoms—those that are present in the TMD itself—has so far received little attention^{11,29,30}. Such self-intercalated TMD compounds may exist as local energy minima in the region of the intercalation phase diagram in which a metal-rich stoichiometry is promoted by growth conditions involving metal atoms at high chemical potential. However, growth windows of TMDs using high metal chemical potentials have so far remained relatively unexplored^{31,32}.

In this work, the growth of 2D TMDs using both molecular beam epitaxy (MBE) and chemical vapour deposition (CVD) methods was investigated under high metal chemical potentials. We discovered that—independent of the growth method used—a metal-rich chemical potential promotes the self-intercalation of a metal (M) into MX, MX₂

¹Department of Materials Science and Engineering, National University of Singapore, Singapore, Singapore. ²Department of Chemistry and Centre for Advanced 2D Materials, National University of Singapore, Singapore, Singapore. ³Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China. ⁴CAMD and Center for Nanostructured Graphene (CNG), Department of Physics, Technical University of Denmark, Kongens Lyngby, Denmark. ⁵School of Materials Science and Engineering, Nanyang Technological University, Singapore, Singapore. ⁶NUSNNI-NanoCore, National University of Singapore, Singapore, Singapore. ⁷School of Physical Sciences and CAS Centre for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing, China. ⁸State Key Laboratory of Optoelectronic Materials and Technologies, Centre for Physical Mechanics and Biophysics, Sun Yat-sen University, Guangzhou, China. ⁹These authors contributed equally: Xiaoxu Zhao, Peng Song. ✉e-mail: luox77@mail.sysu.edu.cn; steve.pennycook@nus.edu.sg; chmlhkp@nus.edu.sg

or M_2X_3 layered 2D compounds (M, metal; X, chalcogen), producing covalently bonded M_xX_y compounds. We term this class of materials ic-2D. Taking TaS_2 as an example, the intercalated Ta atoms occupy the octahedral vacancies in the vdW gap to form distinct topographical patterns, as verified by atomic resolution scanning transmission electron microscopy–annular dark field (STEM–ADF) imaging. By varying the ratio of intercalating atoms to octahedral vacancies in the vdW gap, we grew Ta_xS_y or Ta_xSe_y films and quantified the extent of Ta-intercalation using σ , the percentage of initial total vacancy sites that are occupied by intercalated atoms. Our results indicate that self-intercalation is common to a broad class of vdW crystals, and it offers a powerful approach through which to transform layered 2D materials into ultrathin, covalently bonded ic-2D crystals with ferromagnetic properties.

We first describe the self-intercalation of native atoms—that is, Ta—into a TaS_2 bilayer during MBE deposition on a silicon wafer, as a means to demonstrate the formation of an ic-2D film via octahedral vacancy filling of a 2D bilayer material. Wafer-scale Ta-intercalated TaS_2 bilayer films were grown on 2-inch, 285-nm SiO_2/Si wafers in a dedicated MBE system¹⁴. Ultra-pure Ta and S molecular beams were evaporated from an e-beam evaporator and a sulfur cracker cell equipped with a valve, respectively (Fig. 1a, b). We could routinely grow 2H-phase TaS_2 bilayer films using a high S chemical potential—that is, a Ta-to-S flux ratio of around 1:10 (Fig. 1a, Supplementary Fig. 1)—for 3 h and a substrate temperature of 600 °C. When the Ta:S flux ratio was increased to 1:6 (Fig. 1b, c), the film became non-stoichiometric with respect to TaS_2 owing to the excess of Ta atoms. A fingerprint of the Ta-rich environment is the presence of Ta adatoms (Fig. 1d) occupying the centre of the honeycombs (Fig. 1e) or situated on top of the Ta sites (Fig. 1f) in the monolayer TaS_2 film, as observed by STEM when the growth was interrupted partway through (Supplementary Fig. 2). When Ta and S are continually supplied in the appropriate ratio, the Ta adatoms become embedded in the TaS_2 structure, occupying the octahedral vacancies between two S layers (Fig. 1g). The ic-2D crystals therefore have a sequential, TaS_2 -Ta- TaS_2 -Ta layer-by-layer growth mechanism; as such, multilayer or bulk-phase ic-2D crystals can be readily accessed simply by increasing the growth time. The thermodynamic stability of such intercalated phases was assessed using energy-composition phase diagrams generated through density functional theory (DFT) calculations (Fig. 1h). It was found that stoichiometric H-phase TaS_2 is formed only under S-rich conditions (when the chemical potential of sulfur, μ_S , exceeds -5.3 eV), whereas at higher Ta:S flux ratios (low μ_S), various Ta-intercalated Ta_xS_y configurations—ranging from Ta_9S_{16} (25% Ta intercalation) to Ta_8S_{12} (66.7% Ta intercalation)—entered a thermodynamically stable state.

Notably, a Ta:S flux ratio of approximately 1:6 produced a $\sqrt{3}a \times \sqrt{3}a$ superlattice of Ta atoms (Fig. 2a) sandwiched between two TaS_2 monolayers. The extent of intercalation (σ) was 33.3%, and the overall stoichiometry of the crystal became Ta_7S_{12} , as corroborated by both the real-space STEM image (Fig. 2b) and the corresponding fast Fourier transform (FFT) pattern (Fig. 2c). Image simulation and sequential STEM images capturing the diffusion of intercalated atoms showed that the periodically arranged bright spots in the STEM image were induced by the intercalation of Ta (Fig. 2d, Supplementary Information section 1, Supplementary Videos 1, 2). We also collected STEM cross-section images (Fig. 2e, f) to verify the existence of an intercalated Ta atomic layer in the vdW gap of ic-2D films grown by CVD.

The homogeneous Ta_7S_{12} phase was grown directly on a 2-inch silicon wafer (Supplementary Fig. 3). The Ta_7S_{12} film was formed by the coalescence of nano-domain crystals (around 50 nm) separated by mirror twin boundaries or tilted grain boundaries (Supplementary Information section 2). The amorphous islands and gaps seen in the STEM images were attributed to the poor stability of Ta_xS_y and to sample damage incurred during transfer. Energy dispersive X-ray spectroscopy (EDS) and electron energy loss spectroscopy (Supplementary Fig. 4) verified that the film was composed solely of Ta and S, with no foreign

elements, and X-ray photoelectron spectroscopy (Supplementary Fig. 5) confirmed that the chemical stoichiometry agreed very well with Ta_7S_{12} . The Raman spectra of the film exhibited two prominent E_g^3 and A_g^3 peaks at 300 cm^{-1} and 400 cm^{-1} , respectively, matching those of H-phase TaS_2 films. The fingerprint of the intercalation was a series of minor peaks in the 100 cm^{-1} to 170 cm^{-1} range (Supplementary Fig. 6), which were absent in pure H-phase TaS_2 ³³ and are attributed to the covalent bonds between the intercalated Ta atoms and their octahedrally coordinated S atoms (Supplementary Fig. 7).

25% Ta-intercalated TaS_2 has a stoichiometry of Ta_9S_{16} and was produced at a slightly lower Ta chemical potential than Ta_7S_{12} , corresponding to a Ta:S ratio of around 1:8. The intercalated Ta atoms occupy the octahedral vacancies in every $2a \times \sqrt{3}a$ unit length, and this phase was distinguished by the square symmetry of the intercalated atomic lattice (Fig. 2g, k, Supplementary Fig. 8). When the Ta:S flux ratio was further increased to 1:5, a $Ta_{10}S_{16}$ phase ($\sigma = 50\%$) was successfully grown (Fig. 2h). The intercalation concentration—the percentage of total vacancy sites that were occupied—was determined to be exactly 50% via atom counting (Supplementary Fig. 9). Notably, this phase is characterized by atomic chains that are interconnected over a short range, forming an overall glassy phase. Clear diffusive rings were observed in the proximity of the first-order FFT spots (Fig. 2l, Supplementary Fig. 10), confirming this short-range ordered structure³⁴. When the Ta:S flux ratio was further increased, the glassy phase was retained, but the short atomic chains became denser before fully evolving into a complete atomic plane when σ reached approximately 100% (Supplementary Fig. 11). The use of growth conditions intermediate between those that give rise to high-symmetry phases resulted in phase separations, and atomically sharp domain boundaries separating two high-symmetry phases were apparent (Supplementary Information section 3).

To verify that ic-2D films could be produced by methods other than MBE, we used CVD to grow self-intercalated Ta_xSe_y crystals using excess Ta precursors. The crystal domains of these films were in the micro-metre range—considerably larger than the nanosized domains grown by MBE (Supplementary Fig. 12). A typical Ta_8Se_{12} crystal ($\sigma = 66.7\%$) is depicted in Fig. 2i. Notably, it possesses a Kagome lattice belonging to the P_6 wallpaper symmetry group. A well-defined $\sqrt{3}a \times \sqrt{3}a$ periodic lattice can be unambiguously identified in the atomic-resolution STEM image (Fig. 2m; for the simulated image, see Supplementary Fig. 13). At even higher Ta chemical potential we successfully synthesized Ta_9Se_{12} crystals ($\sigma = 100\%$), in which the trigonal prismatic vacant sites in AA-stacked Ta_9Se_{12} were fully occupied (Fig. 2j)—as seen from the top view (Fig. 2n) and side view (Fig. 2e, Supplementary Fig. 14) STEM images. By precisely controlling the metal:chalcogen ratio during growth, we can prepare a full range of Ta-intercalated Ta_xSe_y or Ta_xS_y compounds with intercalation levels ranging from $\sigma = 25\%$ to over 100%, as verified by EDS (Supplementary Fig. 15, Supplementary Table 1).

In ic-2D films, the intercalated Ta atoms are octahedrally coordinated to the S_6 cage, as opposed to the trigonal-prismatic coordination that is adopted in pristine TaS_2 . Charge transfer from the intercalated Ta atoms to the TaS_2 host layers creates new electron ordering and modifies the Ta d -band splitting. Because the amount of charge transfer is dependent on the concentration of the intercalant, the system can be tuned. To investigate whether ferromagnetic order is present in the intercalated samples, magneto-transport measurements were carried out on MBE-grown Ta_7S_{12} ($\sigma = 33.3\%$) with a predominantly 2H_s stacking registry (Fig. 3a, Supplementary Fig. 16) and bilayer thickness (Supplementary Fig. 17). Figure 3c shows the temperature-dependent resistivity, in which a non-saturating upturn is observed below 30 K owing to the disorder-induced metal–insulator transition in the polycrystalline sample³⁵. Linear magnetoresistance up to 9 T is observed at low temperatures in Ta_7S_{12} (Fig. 3d), owing to density and mobility fluctuations³⁶. The anomalous Hall effect (AHE) arises from the interplay of spin–orbit interactions and ferromagnetic order, and is

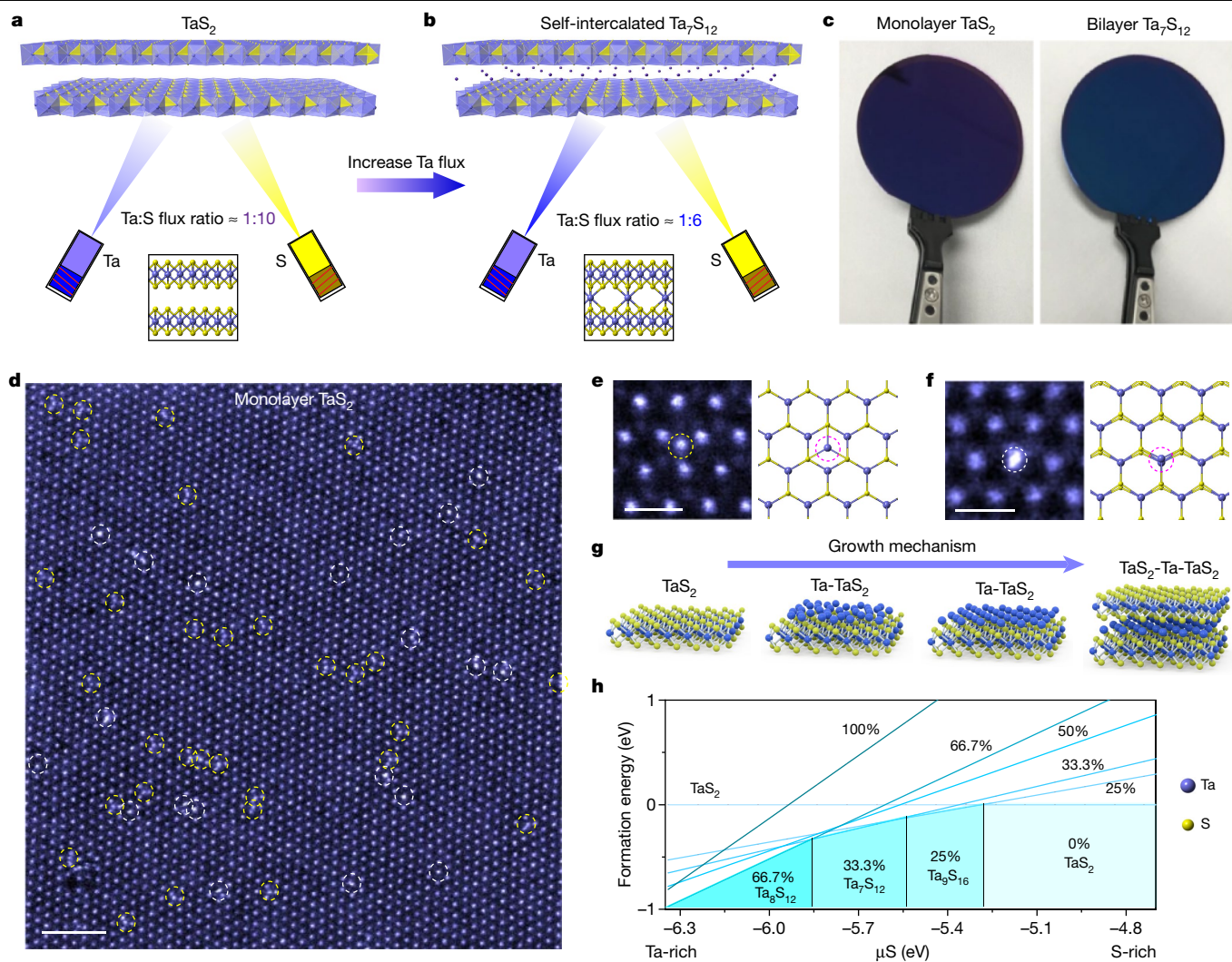


Fig. 1 | Self-intercalation in TaS₂ crystals. **a, b**, Schematic showing the growth of pristine TaS₂ (**a**) and self-intercalated Ta₇S₁₂ (**b**) by MBE under a low and a high Ta-flux environment, respectively. The lower Ta flux produces stoichiometric TaS₂, whereas a higher Ta flux generates a self-intercalated phase. **c**, Photographs of monolayer TaS₂ and bilayer Ta₇S₁₂, grown by MBE on a 2-inch SiO₂/Si wafer. **d–f**, Atomic-resolution STEM-ADF image of monolayer TaS₂ under Ta-rich conditions (**d**), showing an abundance of interstitial Ta atoms at

the centre of honeycomb (**e**) or on top of the Ta site (**f**). In **e, f**, the corresponding atomic models are depicted on the right. **g**, Schematic depicting the layer-by-layer growth of ic-2D crystals. **h**, Calculated formation energies of various self-intercalated Ta_xS_y phases with intercalation concentrations of 25%, 33.3%, 50%, 66.7% and 100%, as a function of the chemical potential of sulfur. Scale bars: **d**, 2 nm; **e, f**, 0.5 nm.

a potentially useful probe of spin polarization. We observed AHE in Ta₇S₁₂ in addition to the linear ordinary Hall effect (OHE). Figure 3e shows a nonlinear Hall effect in the proximity of zero magnetic field and a linear OHE at high field. Although both multiband conduction and the AHE contribute to the nonlinear Hall effect, the observed linear OHE suggests single-carrier (hole) conduction in Ta₇S₁₂ and thus excludes multiband transport as the origin of the nonlinear Hall effect^{37,38}. The nonlinear Hall effect is therefore ascribed to AHE, which arises from ferromagnetism in conductors³⁹. After subtracting the linear OHE, anomalous Hall resistance of up to 0.75 Ω is observed at 1.5 K; this decreases with increasing temperature and disappears at 10 K, which is in line with Monte Carlo simulations based on the Ising model (Supplementary Fig. 18).

The effects of self-intercalation on the electrical properties of TMDs were further assessed in Ta₈Se₁₂ ($\sigma = 66.7\%$), which forms a Kagome lattice. It was found that the intercalation of Ta atoms and the formation of Kagome lattices stabilize the charge-density wave states. The temperature-dependent Hall signal reveals an AHE below 15 K and confirms ferromagnetic order in Ta₈Se₁₂ (Supplementary Fig. 19, 20).

We performed DFT calculations in order to understand the origin of the magnetization in self-intercalated Ta₇S₁₂. Perfect bilayer 2H_s-stacked TaS₂ (Supplementary Fig. 21) possesses a non-magnetic ground state, in which ferromagnetism can be induced by the double exchange mechanism⁴⁰, triggered by the charge transfer from intercalated Ta to pristine TaS₂ (Fig. 3f). When the intercalated Ta adopts a $\sqrt{3}a \times \sqrt{3}a$ superstructure, six S atoms bond with one intercalated Ta atom to form an octahedral unit in the vdW gap. By contrast, each S atom is shared by three Ta atoms in the pristine TaS₂ layer. This difference in local bonding arrangement induces charge transfer from the octahedral-coordinated intercalated Ta atom to the prismatic-coordinated Ta atom in the TaS₂ layer (Fig. 3f). In pristine H-phase TaS₂, the Ta *d* orbitals and the S *p* orbitals are well separated in terms of energy, with the states at the Fermi level having mainly Ta *d*_{z² and Ta *d*_{x²-y² characteristics (Supplementary Fig. 21). In Ta₇S₁₂ ($\sigma = 33.3\%$), the intercalated Ta atoms introduce additional spin-split bands across the Fermi level, and a magnetic ground state develops (Fig. 3g, h). The magnetic moments are localized on the *d* orbitals of the intercalated Ta atom, as evidenced by the calculated intercalated Ta orbital-resolved spin-up and spin-down band structures in Fig. 3g}}

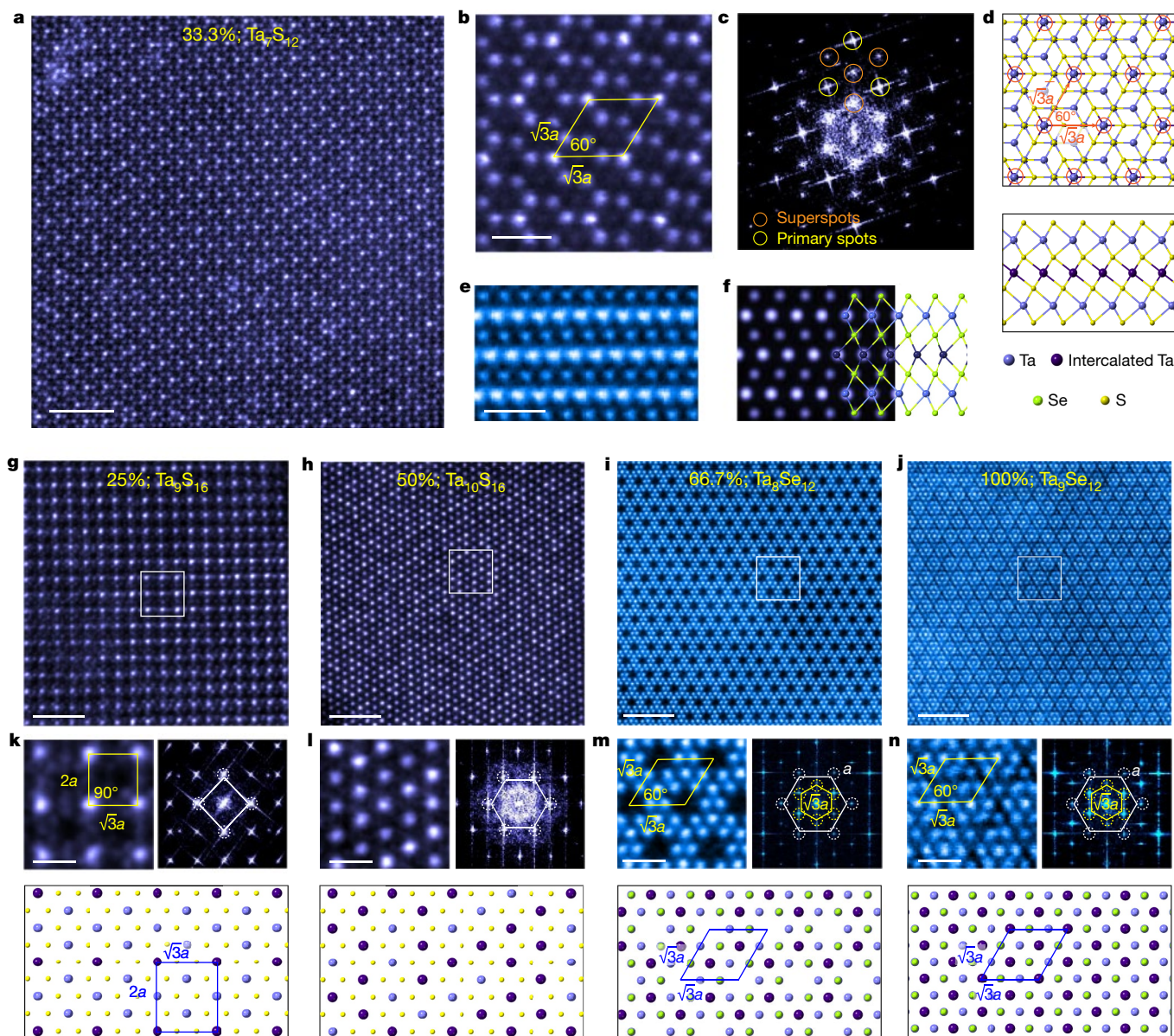


Fig. 2 | Compositional engineering of Ta_xS_y and Ta_xSe_y with different concentrations of intercalated Ta. **a, b**, Atomic-resolution STEM-ADF images of self-intercalated Ta_7S_{12} , grown by MBE, showing the well-defined $\sqrt{3}a \times \sqrt{3}a$ superstructure (**a**), and an enlarged image (**b**). **c**, The corresponding FFT pattern of **a**, with $\sqrt{3}a$ superspots highlighted by orange circles. **d**, Atomic model of self-intercalated Ta_7S_{12} . **e, f**, STEM cross-section view of 100% Ta-intercalated $\text{Ta}_6\text{Se}_{12}$ (**e**) and its corresponding simulated image derived from

the DFT-optimized atomic model (**f**). **g–j**, Atomic-resolution STEM images of 25% Ta-intercalated Ta_9S_{16} (**g**), 50% Ta-intercalated $\text{Ta}_{10}\text{S}_{16}$ (**h**), 66.7% Ta-intercalated $\text{Ta}_8\text{Se}_{12}$ (**i**) and 100% Ta-intercalated $\text{Ta}_6\text{Se}_{12}$ (**j**) ic-2D crystals. **k–n**, Left, enlarged STEM images corresponding to the regions highlighted with white boxes in **g–j**, respectively; right, the corresponding FFT patterns; bottom, the corresponding atomic models. Scale bars: **a, g–j**, 2 nm; **b, e, k–n**, 0.5 nm.

and Fig. 3h, respectively. The states at the Fermi level comprise the prismatic-centred Ta d_{z^2} orbitals hybridized with the spin-up band of the $d_{x^2-y^2}$ orbital of the intercalated Ta. However, only the intercalated Ta atoms exhibit a net spin density, as illustrated in Fig. 3i, in which the top view spin density isosurface matches the shape of the $d_{x^2-y^2}$ orbital. In addition, the non-magnetic $3a \times 3a$ charge-density wave state of Ta_7S_{12} can be ruled out owing to its relative instability compared with the ferromagnetic state⁴¹.

The existence of a magnetic moment correlates with a large degree of charge transfer between the intercalated Ta and the TaS_2 layers. Strong charge transfer occurs when the proportion of intercalated Ta atoms is low, whereas charge transfer becomes relatively weak in a heavily intercalated (Fig. 3j) compound, in accordance with the calculated

charge difference and the variation of Bader charge on the Ta atoms (Supplementary Fig. 22, Supplementary Table 2).

To investigate whether the self-intercalation phenomenon occurred for other TMDs, we performed a high-throughput DFT study of 48 different intercalated TMD bilayers, using a semi-automated workflow for maximal consistency and veracity⁴². Specifically, we considered TMDs of the transition metals Mo, W, Nb, Ta, Ti, Zr, Hf, V, Cr, Mn, Fe, Co, Ni, Pd and Pt, as well as Sn, and the chalcogens S, Se and Te (Fig. 4a) at σ values of 33.3% or 66.7%. Out of this set of TMDs, we observed that 14 bilayer configurations— Ti_8S_{12} , $\text{Ti}_8\text{Se}_{12}$, $\text{Ti}_8\text{Te}_{12}$, Co_7S_{12} , $\text{Co}_7\text{Se}_{12}$, $\text{Co}_7\text{Te}_{12}$, Nb_7S_{12} , $\text{Nb}_7\text{Se}_{12}$, $\text{Nb}_7\text{Te}_{12}$, Mo_7S_{12} , $\text{Mo}_7\text{Se}_{12}$, Ta_7S_{12} , $\text{Ta}_7\text{Se}_{12}$ and $\text{Ta}_7\text{Te}_{12}$ (highlighted by specific σ values and chalcogens in Fig. 4a and Supplementary Table 3 for magnetic moment)—develop ferromagnetic

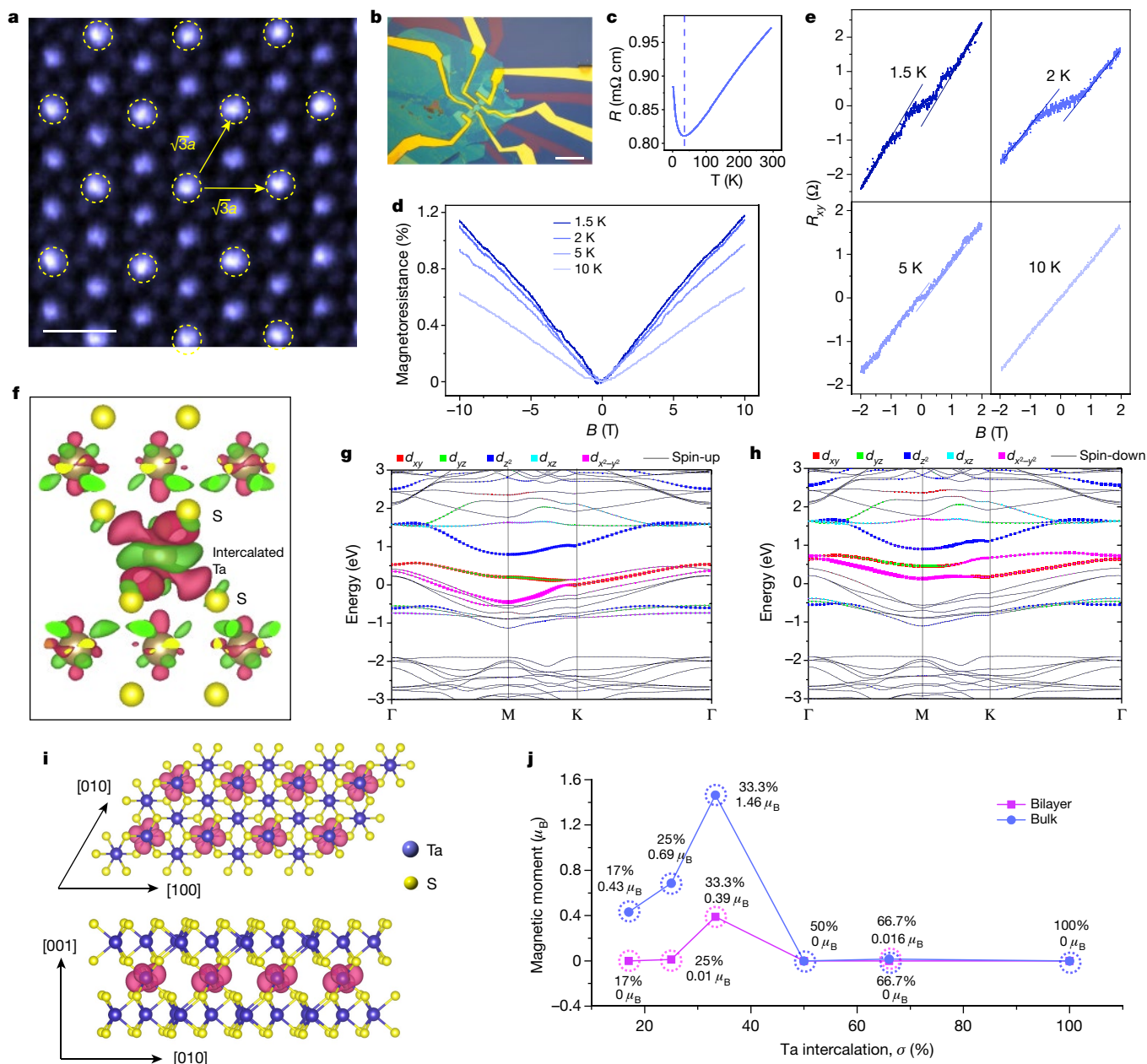


Fig. 3 | Ferromagnetism in Ta-intercalated Ta_7S_{12} ic-2D crystals.

a, Atomic-resolution STEM-ADF image of a typical self-intercalated Ta_7S_{12} film. This image was collected using a half-angle range from about 30 mrad to 110 mrad to enhance the contrast of S. **b**, Optical microscopy image of a Ta_7S_{12} Hall bar device encapsulated with hexagonal boron nitride. **c**, Resistivity of the Ta_7S_{12} ic-2D crystal as a function of temperature. **d**, **e**, Temperature-dependent magnetoresistance (**d**) and Hall resistance (R_{xy}) (**e**) of Ta_7S_{12} under an

out-of-plane magnetic field. **f**, Contour plot of charge density difference in Ta-intercalated Ta_7S_{12} . **g**, **h**, Orbital-resolved spin-up (**g**) and spin-down (**h**) band structures of the intercalated Ta in Ta_7S_{12} . **i**, Top view (top) and side view (bottom) spin density isosurface of Ta-intercalated Ta_7S_{12} . **j**, Calculated magnetic moments as a function of the Ta-intercalation concentration (σ) in $2\text{H}_x\text{S}_{12}$ -stacked nonstoichiometric Ta_7S_{12} . Bohr magneton. Scale bars: **a**, 0.5 nm; **b**, 20 μm .

order upon self-intercalation, whereas their parental MX_2 bilayers are nonferromagnetic. Notably, group V and group VI TMDs exhibit strong ferromagnetism after self-intercalation (Fig. 4b). MX_2 bilayers that are intrinsically ferromagnetic—that is, VX_2 , CrX_2 , MnX_2 and FeX_2 —retain ferromagnetism upon self-intercalation (highlighted by orange triangles in Fig. 4a). Among the 14 self-intercalated 2D ferromagnets that we generated, the formation energies of 12 of these—the two exceptions being MoS_2 and MoSe_2 —were lower than or similar to those of the non-intercalated materials (Supplementary Figs. 23, 24), indicating that self-intercalation is energetically feasible.

To validate our theoretical predictions, we attempted to grow a wide variety of ic-2D materials (Fig. 4a). In this figure, blue triangles

indicate that the self-intercalation can be experimentally realized^{11,12}, whereas grey triangles indicate that intercalation was not successful under our experimental conditions. We succeeded in growing several ic-2D crystals—namely $\text{V}_{11}\text{S}_{16}$ (Fig. 4c, Supplementary Fig. 25), $\text{In}_{11}\text{Se}_{16}$ (Fig. 4d, Supplementary Fig. 26) and Fe_xTe_y (Fig. 4e, Supplementary Fig. 27)—by either CVD or MBE. The topological features and corresponding FFT patterns of these crystals are depicted in Fig. 4f–h. The intercalated $\text{V}_{11}\text{S}_{16}$ has a $2a \times 2a$ superstructure, and the intercalation concentration was estimated at 75% (Fig. 4f). $\text{In}_{11}\text{Se}_{16}$ also showed a $2a \times 2a$ superstructure; however, in this case, the intercalated In atoms reveal a signature honeycomb structure (Fig. 4g). The crystal structure of self-intercalated Fe_xTe_y was complicated—additional Fe atoms were

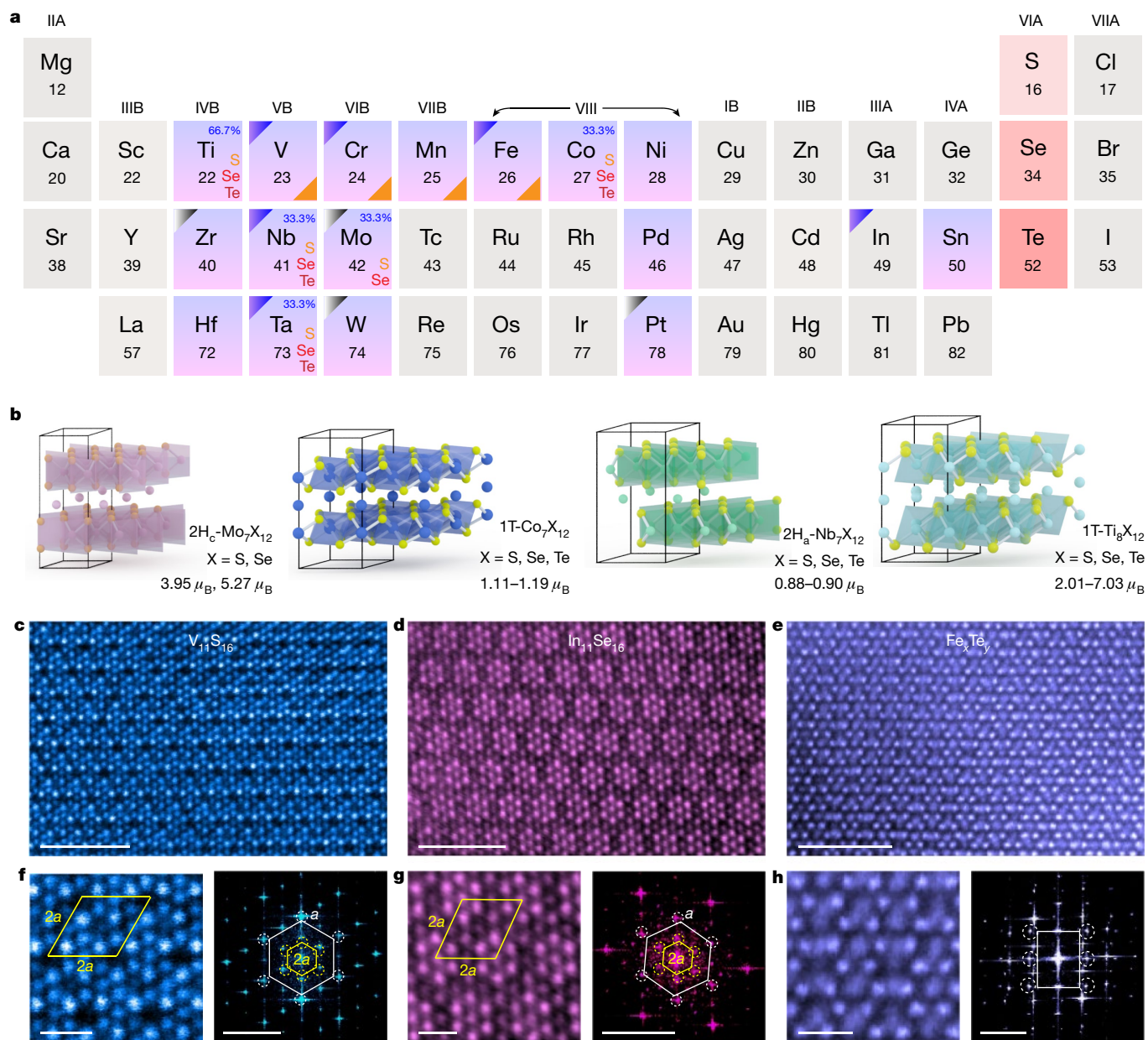


Fig. 4 | A library of ic-2D crystals. a, Periodic table showing metal (blue) and chalcogen (red) combinations that form ic-2D crystals according to our DFT calculations; the list is not exhaustive. Blue triangles indicate that self-intercalation can be experimentally realized, whereas grey triangles indicate that intercalation was not successful under our experimental conditions. MX_2 structures with intrinsic ferromagnetism are highlighted with

orange triangles. **b**, Atomic models, obtained from DFT calculations, of ic-2D crystals that exhibit ferromagnetism. **c–e**, STEM-ADF images of V-intercalated $V_{11}S_{16}$ (**c**), In-intercalated $In_{11}Se_{16}$ (**d**) and Fe-intercalated Fe_xTe_y (**e**). **f–h**, Left, enlarged STEM images of **c–e**, respectively; right, the corresponding FFT patterns. Scale bars: **c–e**, 2 nm; **f–h**, 0.5 nm; FFT patterns in **f–h**, 5 nm^{-1} .

found to be intercalated into the atomic network of the pristine FeTe matrix as interstitials, because telluride-based TMDs offer the largest spacing between the host atoms⁴³. Upon intercalation, the Fe_xTe_y phase reveals new symmetries, as confirmed by the emergence of superspots in the FFT pattern (Fig. 4h). A similar complex intercalation network was also observed in V_xTe_y (Supplementary Fig. 28).

We have developed a robust method to engineer the composition of a broad class of TMDs, by means of self-intercalation with native metal atoms during growth. Because the main principle is the application of high chemical potential of metal atoms to provide the driving force for intercalation during growth, this technique should be compatible with most growth methods. The metal intercalants occupy octahedral vacant sites in the vdW gap, and distinct

stoichiometric phases are produced depending on the levels of intercalation. High-throughput DFT simulations—supported by growth experiments—show that the self-intercalation method is applicable to a large class of 2D layered materials, thus enabling a library of materials with potentially new properties to be created from existing layered materials. Owing to the versatility with which the composition can be controlled, it is possible to tune—in one class of materials—properties such as ferromagnetism and the formation of spin-frustrated Kagome lattices. The implication of this work is that bilayer (or thicker) TMDs can be transformed into ultrathin, covalently bonded 3D materials, with stoichiometry that can be tuned over a broad range by varying the concentration of the intercalants.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2241-9>.

1. Chhowalla, M. et al. The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets. *Nat. Chem.* **5**, 263–275 (2013).
2. Zhou, J. et al. A library of atomically thin metal chalcogenides. *Nature* **556**, 355–359 (2018).
3. Jin, C. et al. Ultrafast dynamics in van der Waals heterostructures. *Nat. Nanotechnol.* **13**, 994–1003 (2018).
4. Wang, C. et al. Monolayer atomic crystal molecular superlattices. *Nature* **555**, 231–236 (2018).
5. Novoselov, K. S. et al. Two-dimensional atomic crystals. *Proc. Natl Acad. Sci. USA* **102**, 10451–10453 (2005).
6. Wan, J. et al. Tuning two-dimensional nanomaterials by intercalation: materials, properties and applications. *Chem. Soc. Rev.* **45**, 6742–6765 (2016).
7. Friend, R. H. & Yoffe, A. D. Electronic properties of intercalation complexes of the transition metal dichalcogenides. *Adv. Phys.* **36**, 1–94 (1987).
8. Wang, X., Shen, X., Wang, Z., Yu, R. & Chen, L. Atomic-scale clarification of structural transition of MoS₂ upon sodium intercalation. *ACS Nano* **8**, 11394–11400 (2014).
9. Tan, S. J. R. et al. Chemical stabilization of 1T' phase transition metal dichalcogenides with giant optical Kerr nonlinearity. *J. Am. Chem. Soc.* **139**, 2504–2511 (2017).
10. Kanetani, K. et al. Ca intercalated bilayer graphene as a thinnest limit of superconducting C₆Ca. *Proc. Natl Acad. Sci. USA* **109**, 19610–19613 (2012).
11. Yang, J. et al. Ultrahigh-current-density niobium disulfide catalysts for hydrogen evolution. *Nat. Mater.* **18**, 1309–1314 (2019).
12. Cui, F. et al. Controlled growth and thickness-dependent conduction-type transition of 2D ferrimagnetic Cr₂S₃ semiconductors. *Adv. Mater.* **32**, 1905896 (2020).
13. Mortazavi, M., Wang, C., Deng, J., Shenoy, V. B. & Medhekar, N. V. Ab initio characterization of layered MoS₂ as anode for sodium-ion batteries. *J. Power Sources* **268**, 279–286 (2014).
14. Fu, D. et al. Molecular beam epitaxy of highly crystalline monolayer molybdenum disulfide on hexagonal boron nitride. *J. Am. Chem. Soc.* **139**, 9392–9400 (2017).
15. Chen, J. et al. Homoepitaxial growth of large-scale highly organized transition metal dichalcogenide patterns. *Adv. Mater.* **30**, 1704674 (2018).
16. Liao, M. et al. Twist angle-dependent conductivities across MoS₂/graphene heterojunctions. *Nat. Commun.* **9**, 4068 (2018).
17. Koski, K. J. et al. Chemical intercalation of zerovalent metals into 2D layered Bi₂Se₃ nanoribbons. *J. Am. Chem. Soc.* **134**, 13773–13779 (2012).
18. Guilmeau, E., Barbier, T., Maignan, A. & Chateigner, D. Thermoelectric anisotropy and texture of intercalated TiS₂. *Appl. Phys. Lett.* **111**, 133903 (2017).
19. Wang, M. et al. Chemical intercalation of heavy metal, semimetal, and semiconductor atoms into 2D layered chalcogenides. *2D Mater.* **5**, 045005 (2018).
20. Dungey, K. E., Curtis, M. D. & Penner-Hahn, J. E. Structural characterization and thermal stability of MoS₂ intercalation compounds. *Chem. Mater.* **10**, 2152–2161 (1998).
21. Gong, Y. et al. Spatially controlled doping of two-dimensional SnS₂ through intercalation for electronics. *Nat. Nanotechnol.* **13**, 294–299 (2018).
22. Chen, Z. et al. Interface confined hydrogen evolution reaction in zero valent metal nanoparticles-intercalated molybdenum disulfide. *Nat. Commun.* **8**, 14548 (2017).
23. Liu, C. et al. Dynamic Ag⁺-intercalation with AgSnSe₂ nano-precipitates in Cl-doped polycrystalline SnSe₂ toward ultra-high thermoelectric performance. *J. Mater. Chem. A* **7**, 9761–9772 (2019).
24. Bouwmeester, H. J. M., van der Lee, A., van Smaalen, S. & Wiegers, G. A. Order–disorder transition in silver-intercalated niobium disulfide compounds. II. Magnetic and electrical properties. *Phys. Rev. B* **43**, 9431–9435 (1991).
25. Wan, C. et al. Flexible n-type thermoelectric materials by organic intercalation of layered transition metal dichalcogenide TiS₂. *Nat. Mater.* **14**, 622–627 (2015).
26. Jeong, S. et al. Tandem intercalation strategy for single-layer nanosheets as an effective alternative to conventional exfoliation processes. *Nat. Commun.* **6**, 5763 (2015).
27. O'Brien, E. S. et al. Single-crystal-to-single-crystal intercalation of a low-bandgap superatomic crystal. *Nat. Chem.* **9**, 1170–1174 (2017).
28. Kumar, P., Skomski, R. & Pushpa, R. Magnetically ordered transition-metal-intercalated WSe₂. *ACS Omega* **2**, 7985–7990 (2017).
29. Kim, S. et al. Interstitial Mo-assisted photovoltaic effect in multilayer MoSe₂ phototransistors. *Adv. Mater.* **30**, 1705542 (2018).
30. Zhang, M. et al. Electron density optimization and the anisotropic thermoelectric properties of Ti self-intercalated Ti_{1+x}S₂ compounds. *ACS Appl. Mater. Interfaces* **10**, 32344–32354 (2018).
31. Wang, S. et al. Shape evolution of monolayer MoS₂ crystals grown by chemical vapor deposition. *Chem. Mater.* **26**, 6371–6379 (2014).
32. Zhao, X. et al. Mo-terminated edge reconstructions in nanoporous molybdenum disulfide film. *Nano Lett.* **18**, 482–490 (2018).
33. Mounet, N. et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).
34. Azizi, A. et al. Spontaneous formation of atomically thin stripes in transition metal dichalcogenide monolayers. *Nano Lett.* **16**, 6982–6987 (2016).
35. Motome, Y., Furukawa, N. & Nagaosa, N. Competing orders and disorder-induced insulator to metal transition in manganites. *Phys. Rev. Lett.* **91**, 167204 (2003).
36. Parish, M. M. & Littlewood, P. B. Non-saturating magnetoresistance in heavily disordered semiconductors. *Nature* **426**, 162–165 (2003).
37. Jiang, Z. et al. Structural and proximity-induced ferromagnetic properties of topological insulator-magnetic insulator heterostructures. *AIP Adv.* **6**, 055809 (2016).
38. Jiang, Z. et al. Independent tuning of electronic properties and induced ferromagnetism in topological insulators with heterostructure approach. *Nano Lett.* **15**, 5835–5840 (2015).
39. Nagaosa, N., Sinova, J., Onoda, S., MacDonald, A. H. & Ong, N. P. Anomalous Hall effect. *Rev. Mod. Phys.* **82**, 1539–1592 (2010).
40. Zener, C. Interaction between the *d* shells in the transition metals. *Phys. Rev.* **81**, 440–444 (1951).
41. Coelho, P. M. et al. Charge density wave state suppresses ferromagnetic ordering in VSe₂ monolayers. *J. Phys. Chem. C* **123**, 14089–14096 (2019).
42. Hastrup, S. et al. The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).
43. Karthikeyan, J., Komsa, H.-P., Batzill, M. & Krasheninnikov, A. V. Which transition metal atoms can be embedded into two-dimensional molybdenum dichalcogenides and add magnetism? *Nano Lett.* **19**, 4581–4587 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Growth of self-intercalated TMD films by MBE

Ta-intercalated Ta_xS_y films were grown in a dedicated MBE chamber (base pressure $<6 \times 10^{-10}$ torr). Before growth, the 2-inch SiO_2 substrates were degassed in the same chamber at 500 °C for 2 h. Ultrapure Ta (99.995%, Goodfellow) and S powders (99.5% Alfa Aesar) were evaporated from a mini electron-beam evaporator and a standard sulfur valved cracker, respectively. The flux density of Ta was precisely controlled by adjusting the flux current. The temperature of the S cracker cell was maintained at 110 °C, and the flux density was controlled by the shutter of the cracker valve. The substrate temperature was maintained at 600–650 °C and the growth time was about 3 h for all thin films. Controlled growth of 25% Ta-intercalated Ta_9S_{16} , 33.3% Ta-intercalated Ta_7S_{12} and 50% Ta-intercalated $\text{Ta}_{10}\text{S}_{16}$ films was achieved when the Ta/S ratio was set at around 1:8, around 1:6 and around 1:5, respectively. A slightly higher growth temperature facilitates the self-intercalation process. After growth, both Ta and S sources were turned off and the sample was further annealed for another 30 min before cooling to room temperature. In-intercalated In_xSe_y samples were grown in a customized MBE chamber (base pressure $<6 \times 10^{-10}$ torr). Before growth, the 1 cm \times 1 cm SiO_2 substrate was degassed in the chamber at 600 °C for 1 h. Ultrapure In_xSe_3 powder (99.99%) and Se pellets (99.999%) were evaporated from a mini electron-beam evaporator and an effusion cell, respectively. The temperature of the Se effusion cell was set at 150 °C with a hot-lip at 220 °C. The substrate temperature was maintained at 400 °C and the growth time was about 2 h. Controlled growth of $\text{In}_{11}\text{Se}_{16}$ films was achieved when the $\text{In}_2\text{Se}_3/\text{Se}$ ratio was set at around 1:3.

Growth of self-intercalated TMD films by CVD

Ta-intercalated Ta_xSe_y crystals were grown by CVD. Before growth, the SiO_2 substrate was sequentially cleaned using water and acetone, followed by 5 min of O_2 plasma. The furnace was purged by 300 standard cubic centimetres (sccm) of Ar gas for 5 min. Se powders and mixed Ta/ TaCl_3 powders were applied as precursors that were located upstream in a one-inch quartz tube. 40 sccm Ar and 10 sccm H_2 was used as a carrier gas. The samples were grown at 800 °C for 30 min. After growth, the sample was cooled down quickly in a continuous stream of Ar. Controlled growth of 66.7% Ta-intercalated $\text{Ta}_8\text{Se}_{12}$ and 100% Ta-intercalated $\text{Ta}_9\text{Se}_{12}$ was achieved when the content of Se powders and mixed Ta/ TaCl_3 powders were 1 g/15 mg, 1.5 mg and 1 g/30 mg/3 mg, respectively. V-intercalated V_xS_y crystals were grown by CVD. Before growth, the SiO_2 substrates were treated by the same method as indicated for the growth of Ta_xSe_y . Two quartz boats containing 0.5 g S and 0.3 g VCl_3 were loaded upstream of the one-inch quartz tube to dispense the precursors. The carrier gas was 40 sccm Ar together with 10 sccm H_2 . The sample was grown at 680 °C for 30 min. After growth, the sample was cooled quickly under the protection of 100 sccm Ar. Fe-intercalated Fe_xTe crystals were grown by CVD. Before growth, the SiO_2 substrates were treated by the same method as indicated for the growth of Ta_xSe_y . Two quartz boats containing Te (>99.997%) and FeCl_2 (>99.9%) were placed upstream of the one-inch quartz tube to dispense the precursors. The sample was grown at 600 °C for 30 min. After growth, the sample was cooled quickly under the protection of 100 sccm Ar.

Sample characterization

X-ray photoelectron spectroscopy was performed using a SPECS XR 50 X-ray Al K α (1,486.6 eV) source with a pass energy of 30 eV. The chamber base pressure was lower than 8×10^{-10} mbar. Raman spectra were collected at room temperature using the confocal WiTec Alpha 300R Raman Microscope (laser excitation, 532 nm).

STEM sample preparation, image characterization and image simulation

The as-grown TMD films were transferred via a poly (methyl methacrylate) (PMMA) method under the protection of graphene. A

continuous graphene film was coated on fresh Ta_7S_{12} film to protect the surface oxidation via a conventional PMMA method. Subsequently, graphene/ Ta_7S_{12} composites were immersed in 1 M KOH solution to detach the PMMA/ Ta_7S_{12} composite from the SiO_2 substrate, followed by rinsing in deionized water. The PMMA/graphene/ Ta_7S_{12} film was then placed onto a Cu quantifoil TEM grid that was precoated with continuous graphene film⁴⁴. The TEM grid was then immersed in acetone to remove the PMMA films. Atomic-resolution STEM-ADF imaging was performed on an aberration-corrected JEOL ARM200F, equipped with a cold field-emission gun and an ASCOR corrector operating at 60 kV. The convergence semiangle of the probe was around 30 mrad. Image simulations were performed with the QSTEM package assuming an aberration-free probe with a probe size of approximately 1 Å. The convergence semiangle of the probe was set at around 30 mrad, and the accelerating voltage was 60 kV in line with the experiments. The collection angle for high-angle annular dark-field imaging was between 81 and 280 mrad and for medium angle annular dark-field imaging was from 30 to 110 mrad. The phonon configurations were set at 30 with defocus value of 0. The STEM-EDS were collected and processed in an Oxford Aztec EDS system.

Device fabrication and measurements

MBE-grown Ta_7S_{12} and CVD-grown $\text{Ta}_8\text{Se}_{12}$ were selected to fabricate Hall-bar devices using e-beam lithography and e-beam evaporation of Ti/Au (2/60 nm). The MBE-grown Ta_7S_{12} film was then etched into Hall-bar geometry using deep reactive-ion etching. The final devices were encapsulated with hexagonal boron nitride flakes using a dry-transfer method in the glovebox (both O_2 and H_2O less than 1 ppm), to avoid the degradation of Ta_7S_{12} and $\text{Ta}_8\text{Se}_{12}$ under ambient conditions. Low-temperature transport measurements were carried out in an Oxford Teslatron system. All resistances were derived from four-terminal measurements using an SR830 lock-in amplifier, with a constant excitation current of 1 μA .

DFT calculations

First-principles calculations based on DFT were implemented in the plane wave code VASP⁴⁵ using the projector-augmented wave potential approach. For the exchange and correlation functional, both the local density approximation and the Perdew–Burke–Ernzerhof (PBE)⁴⁶ flavour of the generalized gradient approximation were used, and no discernible difference were found in the results. A kinetic energy cut off of 500 eV was used for the TaS_2 . A Monkhorst Pack⁴⁷ k -grid sampling with a k -point density of 6.0 \AA^{-1} was used for geometry optimization. For thin-film calculations, a vacuum thickness of 20 Å was added in the slab to minimize the interaction between adjacent image cells. Geometry optimization was performed with the maximum force convergence criterion of $0.005 \text{ eV \AA}^{-1}$. To treat the strong on-site Coulomb interaction of localized Ta d orbitals, we used Dudarev's approach⁴⁸ with an effective U parameter of $U_{\text{eff}} = 3.0 \text{ eV}$. The zone centre phonon modes were calculated using density functional perturbation theory with the local density approximation functionals.

High-throughput DFT calculations

These were carried out with the electronic structure code GPAW⁴⁹ following a semi-automated workflow for maximal consistency and accuracy⁴². The relaxations of the self-intercalated bilayers were done on a Monkhorst-Pack⁴⁷ grid with a k -point density of 6.0 \AA^{-1} using the PBE⁴⁶ and BEEF-vdW functionals⁵⁰ for describing exchange-correlation effects. A vacuum of 15 Å was used in the out-of-plane direction to avoid non-physical periodic interactions. The plane-wave expansion was cut off at 800 eV. All systems were relaxed until the maximum force on any atom was 0.01 eV \AA^{-1} and the maximum stress on the unit cell was $0.002 \text{ eV \AA}^{-3}$. All systems were calculated in the intercalated structure with both a spin-paired calculation and a spin-polarized calculation. If the total energy of the spin-polarized structure was found to be more than

0.01 eV per atom lower than the spin-paired structure, the structure was concluded to be magnetically more stable than its non-magnetic counterpart. The atomic structures of calculated self-intercalated TMDs (33.3% and 66.7% intercalation concentration) are presented in Supplementary Fig. 29, in which the polymorphism of single-layer MoX_2 , WX_2 , NbX_2 and TaX_2 (X = S, Se and Te) reveals an H-phase, whereas the rest of the TMDs are T-phase, adopting an AA stacking polytype. MoX_2 and WX_2 adopt the AA' stacking order whereas NbX_2 and TaX_2 adopt AB' stacking. All intercalants occupy the octahedral vacancies in the vdW gap.

Data availability

The main data supporting the findings of this study are available within the paper and its Supplementary Information. Additional data are available from the corresponding authors upon reasonable request.

Code availability

The Python code is available in the Supplementary Information.

44. Wang, H. et al. High-quality monolayer superconductor NbSe_2 grown by chemical vapour deposition. *Nat. Commun.* **8**, 394 (2017).
45. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab Initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
46. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
47. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
48. Dudarev, S. L. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B* **57**, 1505 (1998).
49. Enkovaara, J. et al. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys. Condens. Matter* **22**, 253202 (2010).

50. Wellendorff, J. et al. Density functionals for surface science: exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **85**, 235149 (2012).

Acknowledgements K.P.L. thanks A*STAR Project ‘Scalable Growth of Ultrathin Ferroelectric Materials for Memory Technologies’ (grant number A1983c0035) and support from the Centre for Advanced 2D Materials, National University of Singapore. W.Z. acknowledges support from the National Key R&D Program of China (2018YFA0305800) and the Natural Science Foundation of China (51622211). S.J.P. is grateful to the National University of Singapore for funding and the Ministry of Education (MOE) for a Tier 2 grant ‘Atomic scale understanding and optimization of defects in 2D materials’ (MOE2017-T2-2-139). Z.L. thanks the MOE for a Tier 2 grant (2017-T2-2-136) and a Tier 3 grant (2018-T3-1-002), and the A*STAR QTE programme. X.L. acknowledges support from the National Natural Science Foundation of China (grant number 11804286) and the Fundamental Research Funds for the Central Universities (grant number 19lgpy263). DFT calculations were performed using resources of the National Supercomputer Center in Guangzhou supported by the Special Program for Applied Research on Super Computation of the NSFC Guangdong Joint Fund (second phase). K.S.T. acknowledges funding from the European Research Council under the European Union’s Horizon 2020 research and innovation programme (grant number 773122, LIMA). The Center for Nanostructured Graphene is sponsored by the Danish National Research Foundation, project DNRF103. We thank J. P. Shi, F. F. Cui and Y. F. Zhang for providing high-quality CVD samples.

Author contributions X.Z., S.J.P. and K.P.L. conceived the idea. S.J.P. and K.P.L. supervised the execution of the whole work. X.Z. and W.Z. performed the electron microscopy experiments and data analysis. X.L., A.C.R.-J. and C.W. performed the DFT calculations and data analysis. A.C.R.-J. and K.S.T. performed the high-throughput DFT calculations. W.F., Y.D., L.K. and Z.L. grew the samples. D.W. and T.V. measured the magnetism. P.S. performed device fabrication and measurement. J.D. and S.N. developed the Python scripts for data analysis. All authors discussed the results and participated in writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2241-9>.

Correspondence and requests for materials should be addressed to X.L., S.J.P. or K.P.L.

Peer review information *Nature* thanks Thomas Heine and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Accelerated discovery of CO₂ electrocatalysts using active machine learning

<https://doi.org/10.1038/s41586-020-2242-8>

Received: 14 April 2018

Accepted: 13 March 2020

Published online: 13 May 2020

 Check for updates

Miao Zhong^{1,2,9}, Kevin Tran^{3,9}, Yimeng Min^{1,9}, Chuanhao Wang^{1,9}, Ziyun Wang¹, Cao-Thang Dinh¹, Phil De Luna^{4,8}, Zongqian Yu³, Armin Sedighian Rasouli¹, Peter Brodersen⁵, Song Sun⁶, Oleksandr Voznyy¹, Chih-Shan Tan¹, Mikhail Askerka¹, Fanglin Che¹, Min Liu¹, Ali Seifitokaldani¹, Yuanjie Pang¹, Shen-Chuan Lo⁷, Alexander Ip¹, Zachary Ulissi³✉ & Edward H. Sargent¹✉

The rapid increase in global energy demand and the need to replace carbon dioxide (CO₂)-emitting fossil fuels with renewable sources have driven interest in chemical storage of intermittent solar and wind energy^{1,2}. Particularly attractive is the electrochemical reduction of CO₂ to chemical feedstocks, which uses both CO₂ and renewable energy^{3–8}. Copper has been the predominant electrocatalyst for this reaction when aiming for more valuable multi-carbon products^{9–16}, and process improvements have been particularly notable when targeting ethylene. However, the energy efficiency and productivity (current density) achieved so far still fall below the values required to produce ethylene at cost-competitive prices. Here we describe Cu-Al electrocatalysts, identified using density functional theory calculations in combination with active machine learning, that efficiently reduce CO₂ to ethylene with the highest Faradaic efficiency reported so far. This Faradaic efficiency of over 80 per cent (compared to about 66 per cent for pure Cu) is achieved at a current density of 400 milliamperes per square centimetre (at 1.5 volts versus a reversible hydrogen electrode) and a cathodic-side (half-cell) ethylene power conversion efficiency of 55 ± 2 per cent at 150 milliamperes per square centimetre. We perform computational studies that suggest that the Cu-Al alloys provide multiple sites and surface orientations with near-optimal CO binding for both efficient and selective CO₂ reduction¹⁷. Furthermore, in situ X-ray absorption measurements reveal that Cu and Al enable a favourable Cu coordination environment that enhances C–C dimerization. These findings illustrate the value of computation and machine learning in guiding the experimental exploration of multi-metallic systems that go beyond the limitations of conventional single-metal electrocatalysts.

To accelerate catalyst discovery, we developed a machine-learning-accelerated, high-throughput density functional theory (DFT) framework¹⁸ to screen materials *ab initio*. We provided this framework with 244 different copper-containing intermetallic crystals from The Materials Project²⁵, from which we enumerated 12,229 surfaces and 228,969 adsorption sites. We then performed DFT simulations on a subset of these sites to calculate their CO adsorption energies (Supplementary Information). These data were used to train a machine learning model, which we used to predict CO adsorption energies on the adsorption sites. The framework then combined the machine-learning-predicted CO adsorption energies with volcano scaling relationships¹⁷ to predict the most catalytically active sites, which have CO adsorption energies (ΔE_{CO}) near to –0.67 eV, a value predicted to produce near-optimal

activity in the volcano scaling relationship (see Supplementary Information and Supplementary Figs. 1, 2 for details on calculating the optimal ΔE_{CO} of –0.67 eV). These optimal sites were simulated using DFT to provide additional training data for the machine learning model. Cycling among DFT simulation, machine learning regression and machine learning prioritization yielded an automated framework that systematically searched for surfaces and adsorption sites with near-optimal CO adsorption energies. In total, the framework carried out about 4,000 DFT simulations, yielding a set of candidates for experimental testing.

Of the candidate materials identified, we found Cu-Al to be the most promising for active and selective CO₂ reduction. We created two-dimensional activity and selectivity volcano plots for CO₂ reduction

¹Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. ²College of Engineering and Applied Sciences, National Laboratory of Solid State

Microstructures, Collaborative Innovation Center of Advanced Microstructure, Nanjing University, Nanjing, China. ³Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.

⁴Materials Science Engineering, University of Toronto, Toronto, Ontario, Canada. ⁵Ontario Centre for Characterization of Advanced Materials (OCCAM), University of Toronto, Toronto, Ontario,

Canada. ⁶National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei, China. ⁷Industrial Technology Research Institute, Material and Chemical Research

Laboratories, Hsinchu, Taiwan. ⁸Present address: National Research Council of Canada, Ottawa, Ontario, Canada. ⁹These authors contributed equally: Miao Zhong, Kevin Tran, Yimeng Min,

Chuanhao Wang. ✉e-mail: zulissi@andrew.cmu.edu; ted.sargent@utoronto.ca

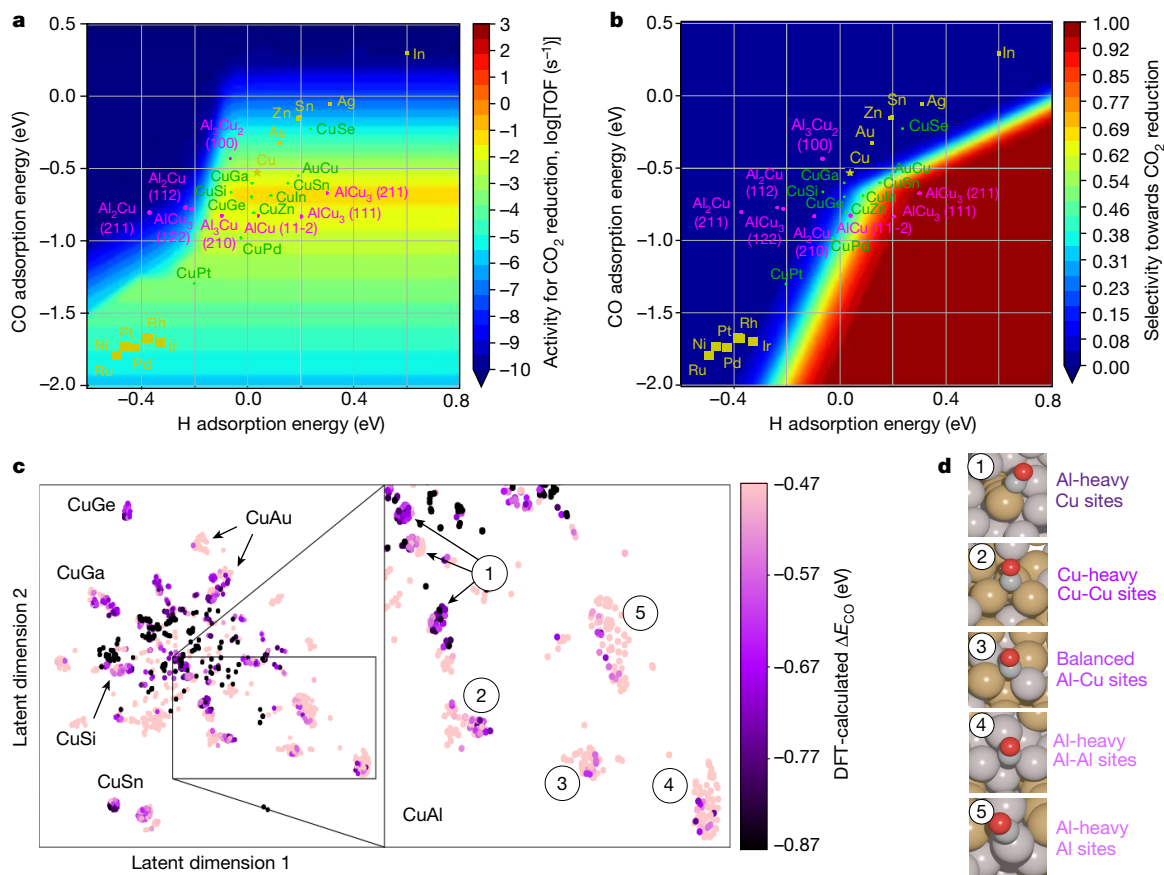


Fig. 1 | Screening of Cu and Cu based compounds using computational methods. **a**, A two-dimensional activity volcano plot for CO₂ reduction. TOF, turnover frequency. **b**, A two-dimensional selectivity volcano plot for CO₂ reduction. CO and H adsorption energies in panels **a** and **b** were calculated using DFT. Yellow data points are average adsorption energies of monometallics; green data points are average adsorption energies of copper alloys; and magenta data points are average, low-coverage adsorption energies

of Cu-Al surfaces. **c**, *t*-SNE¹⁹ representation of approximately 4,000 adsorption sites on which we performed DFT calculations with Cu-containing alloys. The Cu-Al clusters are labelled numerically. **d**, Representative coordination sites for each of the clusters labelled in the *t*-SNE diagram. Each site archetype is labelled by the stoichiometric balance of the surface, that is, Al-heavy, Cu-heavy or balanced, and the binding site of the surface.

(Fig. 1a, b)^{17,26} (Supplementary Information and Supplementary Fig. 3). Figure 1a shows that a CO binding energy near -0.67 eV is required for high activity. It also shows that, given a CO binding energy of about -0.67 eV, a H binding energy above approximately -0.5 eV is required for activity and that a H binding energy above approximately -0.2 eV is required for selectivity towards CO₂ reduction instead of H₂ evolution (Fig. 1a, b).

Since these criteria were met by multiple copper alloy candidates, we pared the list of candidates by visualizing and analysing them in a *t*-SNE diagram¹⁹ (Fig. 1c). Each point on this diagram represents one adsorption site for which we performed a DFT calculation. Points near to one other tend to have similar coordination atoms and surface compositions (Supplementary Information). Clusters of sites represent therefore different adsorption site archetypes (Fig. 1d). Figure 1c shows that Cu-Al exhibits the highest abundance of adsorption sites and site types with near-optimal ΔE_{CO} values, suggesting that Cu-Al alloys may be active across a relatively wide range of surface compositions and site types. The zoomed-in *t*-SNE diagram with example adsorption sites (Fig. 1d) reveals that Al sites tend to bind CO too weakly; Cu sites surrounded by mostly Al atoms may bind CO too strongly; and Cu-Al bridge sites surrounded mostly by Cu atoms are predicted to be active. The low abundance of low ΔE_{CO} sites in Cu-Al alloys also suggests that Cu-Al may be resistant to CO over-binding. We conclude that Cu-Al alloys with a higher Cu content than Al are of potential interest for CO₂ reduction.

To test these hypotheses, we prepared experimentally a suite of Cu-Al model catalysts: ion-implanted Al-on-Cu and evaporated-and-etched Al-on-Cu (see Methods and Supplementary Fig. 4). Each catalyst shows a morphology similar to that of an evaporated pure Cu catalyst (Supplementary Figs. 5–7). Compared with the pure Cu catalyst, which attained a C₂H₄ Faradaic efficiency of 35% at a current density of 600 mA cm⁻² in a 1 M KOH electrolyte in a flow-cell configuration (Supplementary Fig. 8), both ion-implanted and evaporated-and-etched Al-on-Cu catalysts exhibited higher C₂H₄ Faradaic efficiencies of about 60% under the same testing conditions. The CO Faradaic efficiencies on both Cu-Al catalysts were suppressed to about 10%, one-third of that obtained using pure Cu (Supplementary Fig. 9). Incorporating Al on Cu thus increased selectivity towards C₂H₄. Al-on-Cu catalysts maintained about 60% C₂H₄ over 5 h. The Tafel slopes of C₂H₄ production (Supplementary Fig. 9) for pure Cu, ion-implanted, and evaporated-and-etched Al-on-Cu catalysts are 180, 147 and 145 mV per decade, respectively, further highlighting the faster C–C dimerization kinetics with Al-on-Cu catalysts.

To estimate quantitatively the amount of Al incorporated near the Cu surface, we used surface-sensitive Auger electron spectroscopic analysis (Supplementary Figs. 10, 11). This method provides compositional information about the top 1–3 nm of the samples and does so over a relatively large area (100 μm² in our studies)²⁰. We estimated that the molar concentrations of Al on surfaces are 4.5% and 25% for the ion-implanted and evaporated-and-etched Al-on-Cu, respectively. Scanning electron microscopy (SEM) and X-ray spectroscopy analyses

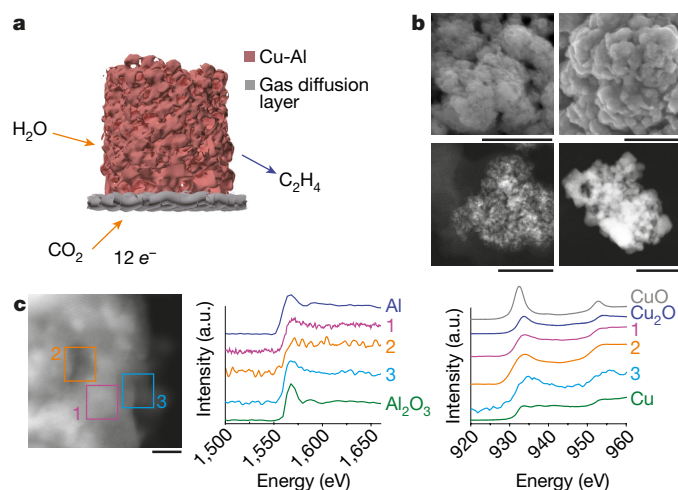


Fig. 2 | Schematic and characterization of de-alloyed Cu-Al catalyst.

a, Schematic of a de-alloyed nanoporous Cu-Al catalyst on a gas diffusion layer for CO₂ electroreduction. **b**, SEM and HAADF-STEM images of de-alloyed Cu-Al catalyst before (left) and after (right) CO₂ electroreduction. The scale bars for the SEM images are 500 nm (top left) and 200 nm (top right). The scale bars for the TEM images are 200 nm (bottom left) and 100 nm (bottom right). **c**, HAADF-STEM image and EELS spectra of the de-alloyed Cu-Al catalyst. Curves numbered 1, 2 and 3 in the EELS spectra represent the EELS results measured at areas 1, 2 and 3 in the corresponding HAADF-STEM image. Al, Al₂O₃, CuO, Cu₂O and Cu EELS results are plotted as references. The scale bar is 5 nm. a.u., arbitrary units.

confirmed no major change of morphologies or Al concentrations for the ion-implanted and evaporated-and-etched Al-on-Cu samples before and after 5 h of CO₂ reduction (Supplementary Figs. 6, 7, 12–15). (See Supplementary Information and Supplementary Figs. 16–19 for detailed operating stability information.) Pourbaix diagrams²¹ (Supplementary Fig. 20) show that both Cu and Al are cathodically protected at potentials more negative than their oxidation potentials of -1.4 V versus a reversible hydrogen electrode (RHE) in a pH 14 electrolyte.

We sought therefore to develop a further optimized Cu-Al catalyst. We explored both thermal evaporation and co-sputtering followed by chemical etching to synthesize de-alloyed nanoporous Cu-Al catalysts (Supplementary Information). As shown in SEM and high-angle angular dark field-scanning transmission electron microscopy (HAADF-STEM) images (Fig. 2b and Supplementary Fig. 21), a nanoporous structure with pore sizes of 5–20 nm was formed. Compared to ion-implanted and evaporated-and-etched Al-on-Cu catalysts, the de-alloyed nanoporous Cu-Al catalysts may offer a higher geometric density of catalytically active sites for adsorption and electroreduction of CO₂. After 5 h of CO₂ electroreduction at a current density of 600 mA cm⁻², the grain size increased, potentially owing to surface reconstruction of Cu and Al in the electrolyte during the reaction (Fig. 2b). Energy-dispersive X-ray spectroscopy analyses in transmission electron microscopy (TEM) and SEM, electron energy loss spectroscopy (EELS) spectra, and elemental mapping in STEM indicated a homogeneous distribution of Al and Cu in de-alloyed catalysts before and also after 5 h of reaction (Fig. 2c and Supplementary Figs. 22–25). We performed HAADF-STEM analysis and found that Cu (111) and (200) facets were observed with interplanar spacings of 0.211 nm and 0.182 nm (Supplementary Fig. 26). Auger electron spectroscopic analysis revealed about 9% Al on the surface following the reaction (Supplementary Figs. 27 and 28).

Given the presence of Cu (111) and (100) surfaces, we used the machine learning model and DFT calculations to analyse how the ratio of Al to Cu on these surfaces affects ΔE_{CO} . First, we enumerated (using Delaunay triangulation²²) the range of adsorption sites on the Cu (111) surfaces having different Al concentrations; and then predicted ΔE_{CO} for

these sites using the machine learning model, thus creating a distribution of ΔE_{CO} values. We repeated this for the Cu (100) surfaces at different Al concentrations. The resulting distributions (Supplementary Fig. 29) show that adding about 12% Al to the Cu (111) surface maximizes the density of sites with ΔE_{CO} values near the optimum of -0.67 eV and that adding 4–12% Al maximizes the density of sites optimal for the Cu (100) surface.

We carried out DFT calculations over the best machine-learning-predicted structures to characterize the changes in reaction energies in the major steps during CO₂ reduction. The reaction energy in the rate-determining step of C–C bond-making¹² decreased from 1.4 eV to 0.6 eV on Cu (111) and from 0.6 eV to 0.4 eV on Cu (100) with the benefit of Al incorporation (Supplementary Figs. 30–33). The DFT results show that the reaction energy of the C–C coupling step (the rate-determining step in the electrochemical CO₂-to-C₂ conversion) is lower for the Cu-Al surfaces compared to that for the corresponding pure Cu surfaces. The DFT results further showed that the reaction energy for forming HO(CH)CH, an intermediate of ethanol²³, was higher than that for forming CCH, an intermediate of C₂H₄ (ref. 23) with Al-containing Cu (Supplementary Fig. 34). Water near the Al atoms may assist the reduction of HOCCH to CCH instead of hydrogenation of HOCCH to HO(CH)CH²³. Thus, the alcohol was suppressed and C₂H₄ production was promoted.

We then systematically evaluated the CO₂ electroreduction performance of the de-alloyed Cu-Al catalysts on carbon-based gas diffusion layer (C-GDL) substrates with about 10% Al at the surfaces at current densities of 200–800 mA cm⁻² in 1 M KOH in flow cells (Fig. 3a and 3b). To quantify the Faradaic efficiencies for each product, we carried out CO₂ electroreduction in the chronopotentiometry mode. As shown in Fig. 3b, we achieved C₂H₄ Faradaic efficiency of 80% at a current density of 600 mA cm⁻². This is a twofold increase compared to the 35% Faradaic efficiency of pure Cu measured under the same conditions. A CO₂-to-C₂H₄ half-cell power conversion efficiency (PCE) in a full-cell CO₂ + H₂O-to-C₂H₄ + O₂ reaction (half-cell C₂H₄ PCE) of 34% was achieved (Fig. 3d), which is similar to the previously published highest half-cell C₂H₄ PCE of about 30% using a plasma-activated copper electrocatalyst¹³ with a C₂H₄ Faradaic efficiency of 60%. This prior work has a much lower current density of around 12 mA cm⁻² in the same electrolyte. An average C₂H₄ Faradaic efficiency of 75 ± 4% was obtained over 17 de-alloyed distinct Cu-Al on C-GDL samples (about 10% Al on the surfaces) at a current density of 600 mA cm⁻². Overall C₂ (multi-carbon product) production Faradaic efficiency was 85–90% when we used the de-alloyed Cu-Al catalyst, higher than the 55–60% using the flat Cu catalyst (Fig. 3c and Supplementary Fig. 9).

We further designed control catalysts—nanoporous Cu on C-GDL with a very low amount of Al on the surface and having similar nanoporosity to that of the de-alloyed Cu-Al catalyst—to clarify the role of morphology (see Methods, Supplementary Information and Supplementary Figs. 35–36). Auger electron spectroscopy analysis revealed that surface Al was a low 2–3% (Supplementary Fig. 37). The C₂H₄ Faradaic efficiency was decreased to 53% at the same current of 600 mA cm⁻² (Fig. 3b and Supplementary Fig. 38).

The Cu-Al on C-GDL catalysts exhibited stable potentials between -1.8 V and -2.1 V versus RHE and a C₂H₄ Faradaic efficiency of 75% over 5 h of continuous operation at 600 mA cm⁻² (Supplementary Fig. 39). After 5 h, the C-GDL gradually lost its hydrophobicity and became flooded with 1 M KOH electrolyte³. CO₂ could therefore no longer diffuse to the catalyst surface for CO₂ reduction.

To improve device stability, we fabricated de-alloyed Cu-Al catalysts on polytetrafluoroethylene (PTFE) substrates whose hydrophobicity is stable over extended operation in a strong alkaline electrolyte³ (Supplementary Information, and Supplementary Figs. 21, 40 and 41). Carbon nanoparticles/graphite were coated on the de-alloyed Cu-Al surface to create a sandwich structure that would distribute the current uniformly over the catalyst to stabilize its surface during

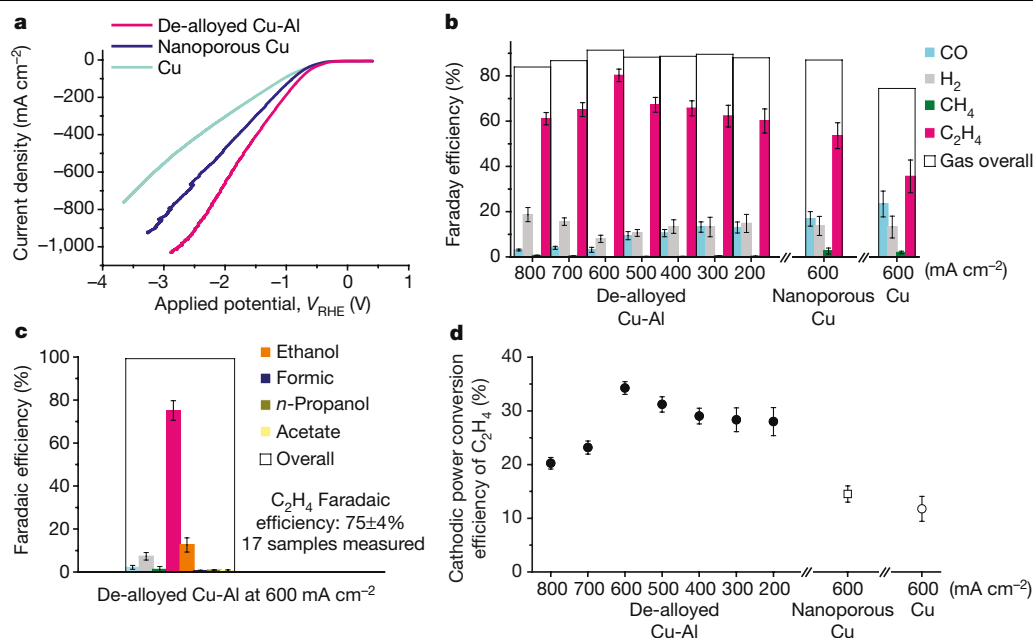


Fig. 3 | CO₂ electroreduction performance on de-alloyed Cu-Al, porous Cu and deposited Cu catalysts on C-GDL substrates in 1 M KOH electrolytes.

a, C₂H₄ production current density versus potential with de-alloyed Cu-Al, nanoporous Cu and evaporated Cu catalysts. **b**, Faradaic efficiencies for gaseous products with de-alloyed Cu-Al catalysts at different applied current densities and with nanoporous Cu and evaporated Cu catalysts at a constant current density of 600 mA cm⁻² obtained using chronopotentiometry. The error bars represent one standard deviation based on five independent

samples measured. **c**, Faradaic efficiencies for all products at an applied current density of 600 mA cm⁻² with 17 de-alloyed Cu-Al samples measured. **d**, Half-cell CO₂-to-C₂H₄ power conversion efficiency with de-alloyed Cu-Al catalysts at different applied current densities and with nanoporous Cu and evaporated Cu catalysts at a constant current density of 600 mA cm⁻² obtained using chronopotentiometry. The error bars represent one standard deviation based on five independent samples measured.

reaction³. As shown in Fig. 4b, c and Supplementary Fig. 42, we achieved C₂H₄ Faradaic efficiencies over 80% in 1 M KOH at a current density of 400 mA cm⁻². Commercial electrolyzers require current densities exceeding 0.2 A cm⁻² for capital costs to be acceptable³³. Compared to the previous best³, we achieved a 2.8× advance in cathodic PCE at 400 mA cm⁻² using Cu-Al. We demonstrated over 100 h of stability at this best condition (Supplementary Figs. 64, 65 and 67).

To improve the overall energy conversion efficiency, we studied Cu-Al performance under different pH conditions²⁷. Experimentally, we found that 3 M KOH (pH 14.5) allowed us to reach 48–52% half-cell C₂H₄ PCE at a current density of 150 mA cm⁻² and was stable over 50 h (Fig. 4b, d). We then optimized the cation concentration by adding an additional 3 M KI into the electrolyte. KI was chosen because the K⁺ cation and I⁻ anion are known to increase CO₂ reduction activity by accelerating the hydrogenation of the key adsorbed CO intermediate^{3,28}. This further diminished the CO Faradaic efficiency to below 0.3% and reduced H₂ production by 3%, increasing the C₂H₄ Faradaic efficiency to 73 ± 4%. As a result, we achieved a 55 ± 2% half-cell C₂H₄ PCE (over ten distinct samples) at 150 mA cm⁻² (Fig. 4b, Supplementary Figs. 43 and 63). Note that the cathodic-side half-cell PCE captures the cathodic CO₂ reduction performance only, and it also does not depend on the location of the reference potential (versus RHE or versus a standard hydrogen electrode, SHE; see the potential diagram in Supplementary Fig. 63). Therefore, the half-cell PCE is useful to compare the energy efficiency on one side of a full-cell reaction^{30–32}. This energy conversion efficiency was stable over 50 h of CO₂ reduction operation. The improved half-cell C₂H₄ PCE in 3 M KOH and 3 M KI electrolytes may benefit from at least one of the following contributions: (1) Al as modulator with Cu to create more active CO₂ reduction sites, (2) the highly nanotextured catalyst surface²⁹, (3) the electrolyte effect from OH⁻, K⁺ and I⁻, all of which are known to increase CO₂ reduction activity^{3,27,28}.

We compare the performance of the de-alloyed Cu-Al/PTFE catalyst with that of the abrupt-interface Cu/PTFE catalyst³ under identical CO₂ electrolysis conditions. The de-alloyed Cu-Al/PTFE catalyst shows

improved Faradaic efficiency and half-cell C₂H₄ PCE under all measured conditions (Supplementary Figs. 64, 65). We note that optimization of electrolysis conditions is crucial to enable Cu-Al to achieve its best CO₂-to-C₂H₄ performance. We also plot the performance of the Cu-Al catalyst compared with that of the previous most efficient abrupt-interface Cu catalyst³ in the reported techno-economic analysis (Supplementary Fig. 66). The Cu-Al catalyst brings the performance into the break-even region; this is an improvement on access to only the below-break-even region in the previous most efficient C₂H₄ electroproduction results.

No obvious leaching of Al and Cu into the solution was observed via inductively coupled plasma atomic emission spectroscopy (ICP-AES) analysis (Supplementary Fig. 44). The concentrations of Cu and Al at time zero are the Cu and Al concentrations in the KOH electrolyte without performing CO₂ electrolysis. Therefore, the detected small amount of Cu and Al in the solutions are impurities from KOH catholyte, which also shows no major change during the reaction, indicating a stable electrolysis system. We further confirmed that the assumed dissolved amounts of Cu and Al from Cu-Al to solution is far below 1% compared to impurity levels in the solution (Supplementary Information).

To investigate the Cu-Al catalyst further, we performed in situ synchrotron X-ray absorption near-edge structure (XANES) analysis under the same testing conditions (see Supplementary Information and Supplementary Fig. 45). Cu-O bonding was observed via both ex situ and in situ XANES analyses with the de-alloyed Cu-Al catalyst before, during and after the reaction²⁴. We used DFT to analyse the reaction energy changes when O is placed on the top surface or in the subsurface of the machine learning-predicted Cu-Al models. The reaction energies in the rate-determining steps in the CO₂ reduction are lower with O in the Cu-Al compared to that of pure Cu (Supplementary Figs. 46–61 and Supplementary Tables 1–8). The XANES spectra of Al in the Cu-Al sample before and after the reaction are shown in Supplementary Fig. 62.

To conclude, we have developed a Cu-Al catalyst for active and selective CO₂ electroreduction to C₂H₄. We demonstrate the

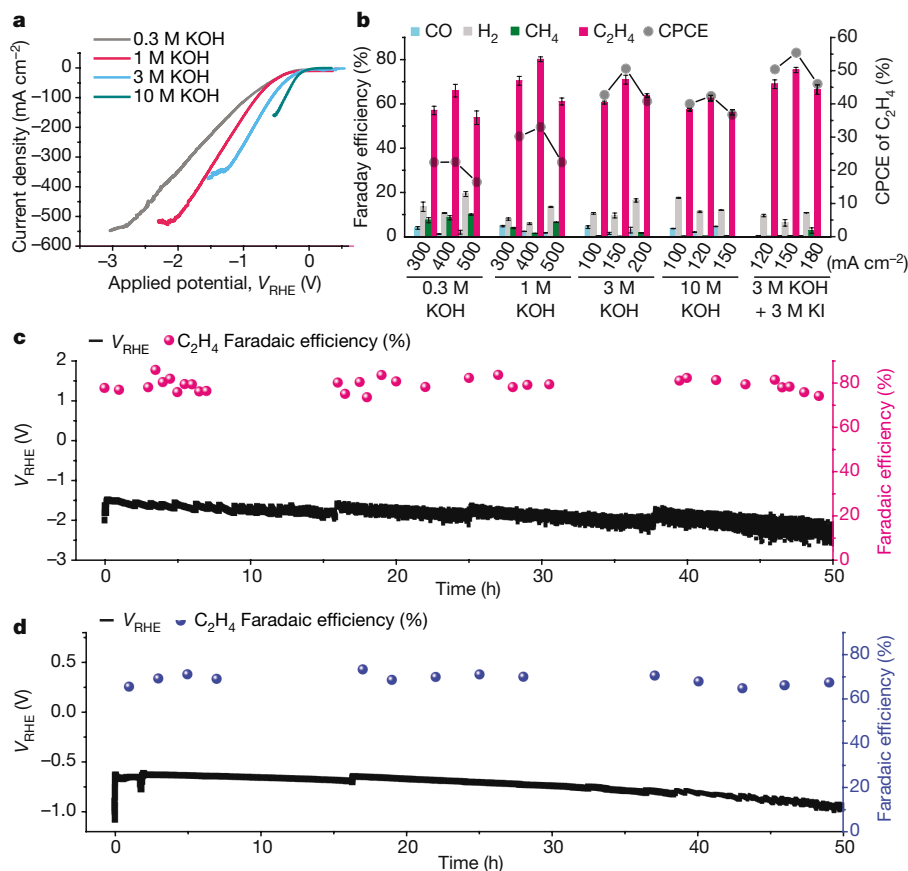


Fig. 4 | CO₂ electroreduction performance on de-alloyed Cu-Al catalysts on PTFE substrates in alkaline electrolytes at different pH values. **a**, C₂H₄ production current density versus potential with de-alloyed Cu-Al in 0.3 M, 1 M, 3 M and 10 M KOH electrolytes. **b**, Faradaic efficiencies for gaseous products with its corresponding C₂H₄ power conversion efficiencies of the de-alloyed Cu-Al catalysts in the different electrolytes and at different applied current densities. The error bars for Faradaic efficiencies measured in 0.3 M and 10 M electrolytes represent one standard deviation based on five independent samples measured. The error bars for Faradaic efficiencies measured in 1 M KOH, 3 M KOH and 3 M KOH + 3 M KI electrolytes represent one standard deviation based on ten independent samples measured. **c**, The CO₂ electroreduction stability of the carbon nanoparticles/de-alloyed Cu-Al/PTFE electrode in a 1 M KOH electrolyte at an applied current density of 400 mA cm⁻². The left axis shows potential (versus RHE; V) versus time (s); the right axis

shows C₂H₄ Faradaic efficiency (%) versus time (s). **d**, The CO₂ electroreduction stability of the carbon nanoparticles/de-alloyed Cu-Al/PTFE electrode in a 3 M KOH electrolyte at an applied current density of 150 mA cm⁻². The left axis shows potential (versus RHE; V) versus time (s); the right axis shows C₂H₄ Faradaic efficiency (%) versus time (s). Note that we passed a small amount of 1 M KI catholyte (pH 5.5–6.5) as a buffer electrolyte before passing the KOH catholyte to protect the Cu-Al catalyst from any possible dissolution into the KOH catholyte. The small amount of KI was then pumped out of the flow-cell system after use as a buffer electrolyte. We convert the potential to V_{RHE} using the equation: $V_{\text{RHE}} = V_{\text{Ag/AgCl}} + 0.199 + 0.059 \times \text{pH}$, in which we use the testing KOH catholyte pH values for calculation. The potentials at time 0 in panels **c** and **d** should be approximately –0.5 V more cathodic. CPCE, cathodic power conversion efficiency.

prediction of promising electrocatalysts by combining volcano relationships, DFT and active machine learning to optimize catalyst performance. The findings suggest avenues towards multi-metal catalysts that outperform single-component catalysts by using an intermediate-binding-optimization and reaction-electrolyte-optimization strategy for multi-carbon production via CO₂ electroreduction.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2242-8>.

- Lin, S. et al. Covalent organic frameworks comprising cobalt porphyrins for catalytic CO₂ reduction in water. *Science* **349**, 1208–1213 (2015).
- Schreier, M. et al. Solar conversion of CO₂ to CO using Earth-abundant electrocatalysts prepared by atomic layer modification of CuO. *Nat. Energy* **2**, 17087 (2017).

- Dinh, C. et al. Sustained high-selectivity CO₂ electroreduction to ethylene via hydroxide-mediated catalysis at an abrupt reaction interface. *Science* **360**, 783–787 (2018).
- Li, C. et al. Electroreduction of carbon monoxide to liquid fuel on oxide-derived nanocrystalline copper. *Nature* **508**, 504–507 (2014).
- Lu, Q. et al. A selective and efficient electrocatalyst for carbon dioxide reduction. *Nat. Commun.* **5**, 3242 (2014).
- Gao, S. et al. Partially oxidized atomic cobalt layers for carbon dioxide electroreduction to liquid fuel. *Nature* **529**, 68–71 (2016).
- Li, Y. C. et al. Electrolysis of CO₂ to syngas in bipolar membrane-based electrochemical cells. *ACS Energy Lett.* **1**, 1149–1153 (2016).
- Jeanty, P. et al. Upscaling and continuous operation of electrochemical CO₂ to CO conversion in aqueous solutions on silver gas diffusion electrodes. *J. CO₂ Util.* **24**, 454–462 (2018).
- Hori, Y. et al. Selective formation of C₂ compounds from electrochemical reduction of CO₂ at a series of copper single crystal electrodes. *J. Phys. Chem. B* **106**, 15–17 (2002).
- Yano, H. et al. Selective electrochemical reduction of CO₂ to ethylene at a three-phase interface on copper(I) halide-confined Cu-mesh electrodes in acidic solutions of potassium halides. *J. Electroanal. Chem.* **565**, 287–293 (2004).
- Peterson, A. A. et al. How copper catalyzes the electroreduction of carbon dioxide into hydrocarbon fuels. *Energy Environ. Sci.* **3**, 1311–1315 (2010).
- Kortlever, R. et al. Catalysts and reaction pathways for the electrochemical reduction of carbon dioxide. *J. Phys. Chem. Lett.* **6**, 4073–4082 (2015).
- Mistry, H. et al. Highly selective plasma-activated copper catalysts for carbon dioxide reduction to ethylene. *Nat. Commun.* **7**, 12123 (2016).

14. Li, Y. et al. Structure-sensitive CO₂ electroreduction to hydrocarbons on ultrathin 5-fold twinned copper nanowires. *Nano Lett.* **17**, 1312–1317 (2017).
15. Lum, Y. et al. Optimizing C–C coupling on oxide-derived copper catalysts for electrochemical CO₂ reduction. *J. Phys. Chem. C* **121**, 14191–14203 (2017).
16. De Luna, P. et al. Catalyst electro-redeposition controls morphology and oxidation state for selective carbon dioxide reduction. *Nat. Catal.* **1**, 103–110 (2018).
17. Liu, X. et al. Understanding trends in electrochemical carbon dioxide reduction rates. *Nat. Commun.* **8**, 15438 (2017).
18. Tran, K. et al. Active learning across intermetallics to guide discovery of electrocatalysts for CO₂ reduction and H₂ evolution. *Nat. Catal.* **1**, 696–703 (2018).
19. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
20. Davis, L. E. et al. *Handbook of Auger Electron Spectroscopy* 2nd edn (Physical Electronics Industries, 1976).
21. Persson, K. A. et al. Prediction of solid-aqueous equilibria: scheme to combine first-principles calculations of solids with experimental aqueous states. *Phys. Rev. B* **85**, 235438 (2012).
22. Montoya, J. H. et al. A high-throughput framework for determining adsorption energies on solid surfaces. *npj Comput. Mater.* **3**, 14 (2017).
23. Xiao, H. et al. Atomistic mechanisms underlying selectivities in C1 and C2 products from electrochemical reduction of CO on Cu (111). *J. Am. Chem. Soc.* **139**, 130–136 (2017).
24. Xiao, H. et al. Cu metal embedded in oxidized matrix catalyst to promote CO₂ activation and CO dimerization for electrochemical reduction of CO₂. *Proc. Natl Acad. Sci. USA* **114**, 6685–6688 (2017).
25. Jain, A. et al. The Materials Project: a materials genome approach to accelerated materials innovation. *APL Mater.* **1**, 011002 (2013).
26. Nørskov, J. K. et al. Trends in the exchange current for hydrogen evolution. *J. Electrochem. Soc.* **152**, J23–J26 (2005).
27. Wang, L. et al. Electrochemical carbon monoxide reduction on polycrystalline copper: effects of potential, pressure, and pH on selectivity toward multicarbon and oxygenated products. *ACS Catal.* **8**, 7445–7454 (2018).
28. Liu, M. et al. Enhanced electrocatalytic CO₂ reduction via field-induced reagent concentration. *Nature* **537**, 382–386 (2016).
29. Zeng, Z. et al. Stabilization of ultrathin (hydroxy)oxide films on transition metal substrates for electrochemical energy conversion. *Nat. Energy* **2**, 17070 (2017).
30. She, Z. W. et al. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **355**, 146 (2017).
31. Larrazábal, G. O. et al. Building blocks for high performance in electrocatalytic CO₂ reduction: materials, optimization strategies, and device engineering. *J. Phys. Chem. Lett.* **8**, 3933–3944 (2017).
32. Whipple, D. T. et al. Prospects of CO₂ utilization via direct heterogeneous electrochemical reduction. *J. Phys. Chem. Lett.* **1**, 3451–3458 (2010).
33. De Luna, P. et al. What would it take for renewably powered electrosynthesis to displace petrochemical processes? *Science* **364**, eaav3506 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Data availability

Source data to generate figures and tables are available from the corresponding authors on reasonable request.

Code availability

Code to generate figures and tables is available from the corresponding authors on reasonable request.

Acknowledgements This work was supported financially by the Ontario Research Fund Research-Excellence Program, the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institute for Advanced Research (CIFAR) Bio-Inspired Solar Energy programme, the University of Toronto Connaught programme, and TOTAL American Services. M.Z. thanks the National Natural Science Foundation of China (grant number 91963121), and S. Tolbert from the University of California, Los Angeles for discussions of de-alloying. S.S. thanks the National Key Research and Development Program of China (grant number 2016YFB0700205) and the National Natural Science Foundation of China (grant number U1632273). We thank R. Morris and D. Sinton from the University of Toronto for discussions. We thank C. McCallum, R. Wolowiec, D. Kopilovic, S. Boccia, A. Ip, M. Liu, Y. Pang, M. Askerka, A. Seifitokaldani, T. T. Zhuang and Z. Liang from the University of Toronto, Canada and C.-W. Huang, L.-J. Chen from National Tsing Hua University, Taiwan, for their help during the course of study. We thank the beamline scientists from the Source optimisée de lumière d'énergie intermédiaire du LURE (SOLEIL) Synchrotron in France for performing in situ X-ray

absorption analyses. This research used resources of the National Energy Research Scientific Computing Center, a Department of Energy (DOE) Office of Science User Facility supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231. Computations were performed on the Southern Ontario Smart Computing Innovation Platform (SOSCIP) Consortium's Blue Gene/Q computing platform. SOSCIP is funded by the Federal Economic Development Agency of Southern Ontario, the Province of Ontario, IBM Canada Ltd, Ontario Centres of Excellence, MITACS and 15 Ontario academic member institutions.

Author contributions E.H.S. supervised the project. M.Z. and E.H.S. conceived the idea. M.Z. and C.W. designed and carried out the experiments. K.T., Z.Y. and Z.U. performed the machine learning studies. K.T., Z.Y., Z.U., Y.M., Z.W. O.V., P.D.L., M.A., M.Z. and E.H.S. discussed the machine learning results. Y.M., Z.W. O.V., P.D.L., M.A., A.S., F.C., K.T., Z.Y. and Z.U. carried out the DFT simulations. P.B. carried out the Auger electron spectroscopy analyses. S.S. and P.D.L. performed X-ray absorption spectroscopy measurements. C.-S.T. and S.-C.L. carried out the TEM analyses. C.-T.D., A.S.R., C.-S.T., M.A., M.L., A.S., Y.P. and A.I. contributed to the discussion of the results. All authors discussed the results and assisted during manuscript preparation.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2242-8>.

Correspondence and requests for materials should be addressed to Z.U. or E.H.S.

Peer review information *Nature* thanks Hailiang Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Rapid growth of new atmospheric particles by nitric acid and ammonia condensation

<https://doi.org/10.1038/s41586-020-2270-4>

Received: 26 September 2019

Accepted: 17 March 2020

Published online: 13 May 2020

Open access

 Check for updates

A list of authors and their affiliations appears at the end of the paper

New-particle formation is a major contributor to urban smog^{1,2}, but how it occurs in cities is often puzzling³. If the growth rates of urban particles are similar to those found in cleaner environments (1–10 nanometres per hour), then existing understanding suggests that new urban particles should be rapidly scavenged by the high concentration of pre-existing particles. Here we show, through experiments performed under atmospheric conditions in the CLOUD chamber at CERN, that below about +5 degrees Celsius, nitric acid and ammonia vapours can condense onto freshly nucleated particles as small as a few nanometres in diameter. Moreover, when it is cold enough (below –15 degrees Celsius), nitric acid and ammonia can nucleate directly through an acid–base stabilization mechanism to form ammonium nitrate particles. Given that these vapours are often one thousand times more abundant than sulfuric acid, the resulting particle growth rates can be extremely high, reaching well above 100 nanometres per hour. However, these high growth rates require the gas-particle ammonium nitrate system to be out of equilibrium in order to sustain gas-phase supersaturations. In view of the strong temperature dependence that we measure for the gas-phase supersaturations, we expect such transient conditions to occur in inhomogeneous urban settings, especially in wintertime, driven by vertical mixing and by strong local sources such as traffic. Even though rapid growth from nitric acid and ammonia condensation may last for only a few minutes, it is nonetheless fast enough to shepherd freshly nucleated particles through the smallest size range where they are most vulnerable to scavenging loss, thus greatly increasing their survival probability. We also expect nitric acid and ammonia nucleation and rapid growth to be important in the relatively clean and cold upper free troposphere, where ammonia can be convected from the continental boundary layer and nitric acid is abundant from electrical storms^{4,5}.

The formation of new particles may mask up to half of the radiative forcing caused since the industrial revolution by carbon dioxide and other long-lived greenhouse gases⁶. Present-day particle formation is thought to predominantly involve sulfuric acid vapours globally^{7–9}. Subsequent particle growth is richer, often involving organic molecules¹⁰. Often growth is the limiting step for the survival of particles from freshly nucleated clusters to diameters of 50 or 100 nm, where they become large enough to directly scatter light and also to seed cloud formation^{11,12}.

New-particle formation in megacities is especially important², in part because air pollution in megacities constitutes a public health crisis¹³, but also because the regional climate forcing associated with megacity urban haze can be large¹⁴. However, new-particle formation in highly polluted megacities is often perplexing, because the apparent particle growth rates are only modestly faster (by a factor of roughly three) than growth rates in remote areas, whereas the vapour condensation sink (to background particles) is up to two orders of magnitude larger (Extended Data Fig. 1). This implies a very low survival probability in the ‘valley of death’, where particles with diameters (d_p) of 10 nm or less have high Brownian diffusivities and will be lost by coagulation scavenging unless they grow rapidly^{7,15}.

Ammonium nitrate has long been recognized as an important yet semivolatile constituent of atmospheric aerosols¹⁶. Especially in

winter and in agricultural areas, particulate nitrate can be a substantial air-quality problem¹⁷. However, the partitioning of nitric acid and ammonia vapours with particulate ammonium nitrate is thought to rapidly reach an equilibrium, often favouring the gas phase when it is warm.

Because ammonium nitrate is semivolatile, nitric acid has not been thought to play an important role in new-particle formation and growth, where very low vapour pressures are required for constituents to be important. Such constituents would include sulfuric acid¹⁸ but also very low vapour pressure organics^{19,20} and iodine oxides²¹. However, it is saturation ratio and not vapour pressure per se that determines the thermodynamic driving force for condensation, and nitric acid can be three or four orders of magnitude more abundant than sulfuric acid in urban environments. Thus, even a small fractional supersaturation of nitric acid and ammonia vapours with respect to ammonium nitrate has the potential to drive very rapid particle growth, carrying very small, freshly nucleated particles through the valley of death in a few minutes. These rapid growth events can exceed 100 nm h^{–1} under urban conditions—an order of magnitude higher than previous observations—and the growth will continue until the vapours are exhausted and conditions return to equilibrium. Such transients will be difficult to identify in inhomogeneous urban environments, yet have the potential

to explain the puzzling observations of new-particle formation in highly polluted megacities.

Nucleation measurements in CLOUD at CERN

Here we report experiments performed with mixtures of nitric acid, sulfuric acid and ammonia vapours under atmospheric conditions in the CERN CLOUD chamber (Cosmics Leaving Outdoor Droplets²²; see Methods for experimental details) from 21 September to 7 December 2018 (CLOUD 13). We varied the temperature from +20 °C to −10 °C, in one case cooling progressively from −15 °C to −25 °C. We adjusted levels of sulfuric acid (H₂SO₄), ammonia (NH₃) and nitric acid (HNO₃), as well as aromatic precursors, to span the ranges typical of polluted megacities. In Fig. 1 we show two representative events at −10 °C. For Fig. 1a we oxidized SO₂ with OH to form H₂SO₄ in the presence of 1,915 parts per trillion volume (pptv) ammonia. The resulting ‘banana’ is typical of such experiments and of ambient observations under relatively clean conditions, with a single nucleation mode that appears shortly after the onset of nucleation and grows at roughly 20 nm h^{−1}. In Fig. 1b we repeated this experiment but also with 5.8 parts per billion volume (ppbv) NO₂, which was oxidized by OH to produce 24 pptv of HNO₃ vapour. The resulting size distribution initially resembles the first case, but when the particles reach about 5 nm, their growth rate accelerates to roughly 45 nm h^{−1}. This activation is reminiscent of cloud-droplet activation and thus suggestive of ‘nano-Köhler’ behaviour and the Kelvin curvature effect²³.

We repeated these experiments over a range of conditions, either forming HNO₃ from NO₂ oxidation or injecting it directly into the CLOUD chamber from an ultrapure evaporation source. We observed this activation and rapid growth behaviour consistently. In Fig. 1c we show the resulting rapid growth rates after activation at −10 °C (green) and +5 °C (purple), plotted against the product of the measured gas-phase HNO₃ and NH₃ mixing ratios. Growth rates are based on the 50% appearance time—the time at which particle number concentrations in each size bin of the rapid growth regime reach 50% of their maximum. Both a strong correlation and a clear temperature dependence are evident; when it is colder, the particles grow at the same rate for a much lower product of vapour concentrations. This is consistent with semivolatile uptake of both species, rate limited by the formation of ammonium nitrate.

To confirm this, we measured the composition of the particles using a filter inlet for gases and aerosols (FIGAERO) iodide (I[−]) chemical ionization mass spectrometer (CIMS), along with the gas-phase vapour concentrations via several CIMS methods. In Fig. 2 we show another rapid growth event, this one at +5 °C (indicated in Fig. 1c with a black outlined purple square). We started with an almost perfectly clean chamber and only vapours present (SO₂, HNO₃ and NH₃) at constant levels (Fig. 2a). Here we injected the HNO₃ without photochemical production so we could independently control HNO₃ and sulfuric acid. The FIGAERO showed no measurable signal in the absence of particles, indicating negligible crosstalk from vapours. We then turned on ultraviolet lights in order to form OH radicals and to initiate SO₂ oxidation to H₂SO₄. Fig. 2b shows the resulting number distribution; as in Fig. 1b, particles appear, grow slowly, and then activate and grow at 700 nm h^{−1}. We again show the 50% appearance time of both modes. In Fig. 2c we show the associated volume distribution. Within 15 min of the onset of particle formation, the volume is dominated by the upper mode near 200 nm. Finally, in Fig. 2d we show a FIGAERO thermogram (signal versus desorption temperature) for particles collected between 10 min and 40 min after the onset of photochemistry. Their composition is dominated by nitrate, with a much smaller but notable sulfate contribution; the semivolatile nitrate desorbs at a much lower temperature than the sulfate. The I[−] chemical ionization is not sensitive to NH₃, but both nitrate and sulfate exist presumably as ammonium salts in the particles.

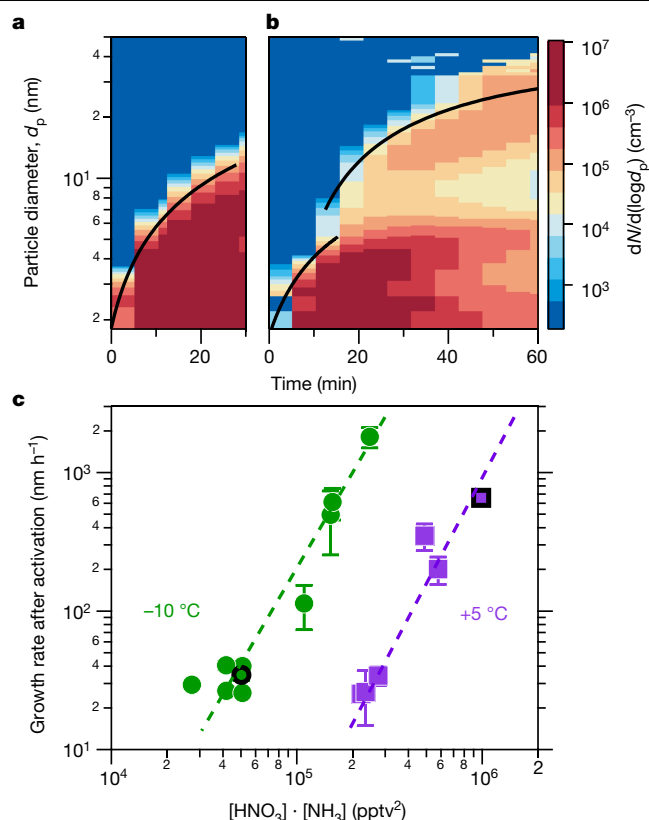


Fig. 1 | Rapid growth events observed in the CERN CLOUD chamber.

a, Particle nucleation and growth (particle growth rate, dd_p/dt) at −10 °C from a mixture of 0.44 pptv sulfuric acid and 1,915 pptv ammonia at 60% relative humidity. Particles form and grow to roughly 10 nm in 30 min. The black curve shows the linear fit to the 50% appearance times. **b**, Particle formation and growth under identical conditions to those in **a**, but with the addition of 24 pptv of nitric acid vapour formed via NO₂ oxidation. Once particles reach roughly 5 nm, they experience rapid growth to much larger sizes, reaching more than 30 nm in 45 min. **c**, Observed growth rates after activation versus the product of measured nitric acid and ammonia levels at +5 °C and −10 °C. The point corresponding to the rapid growth regime for $d_p > 6$ nm in **b** is a black-outlined green circle, and the point corresponding to Fig. 2 is a black-outlined purple square. Growth rates at a given vapour product are substantially faster at −10 °C than at +5 °C, consistent with semivolatile condensation that is rate limited by ammonium nitrate formation. Error bars are 95% confidence limits on the fitting coefficients used to determine growth rates. The overall systematic scale uncertainties of ±10% on the NH₃ mixing ratio and ±25% on the HNO₃ mixing ratio are not shown.

In addition to the correlation of activated particle growth rates with the product of HNO₃ and NH₃ at a given temperature, the observed activation diameter (d_{act}) shows a strong dependence on this product. The activation diameter is evident as a clear kink in the 50% appearance curve, as well as a notable absence of particles in the slower-growth mode above d_{act} . In Extended Data Fig. 2a we show an example of how we determine d_{act} using the emergence of a bimodal size distribution as the defining feature. In Fig. 3a we plot the observed activation diameter at each temperature in a phase space, with [HNO₃] on the log x axis and [NH₃] on the log y axis (both in pptv). The number within each symbol is the observed activation diameter for that experiment. We show the saturation ratio (*S*) of ammonium nitrate at each temperature via a series of diagonal lines in this log–log space (slope = −1); specifically, we show *S* = 1, 5 and 25, emphasizing *S* = 1 as a thick solid line. We also indicate 1:1 [HNO₃]:[NH₃] with a dashed grey line (slope +1); points to the upper left (most of the values) are

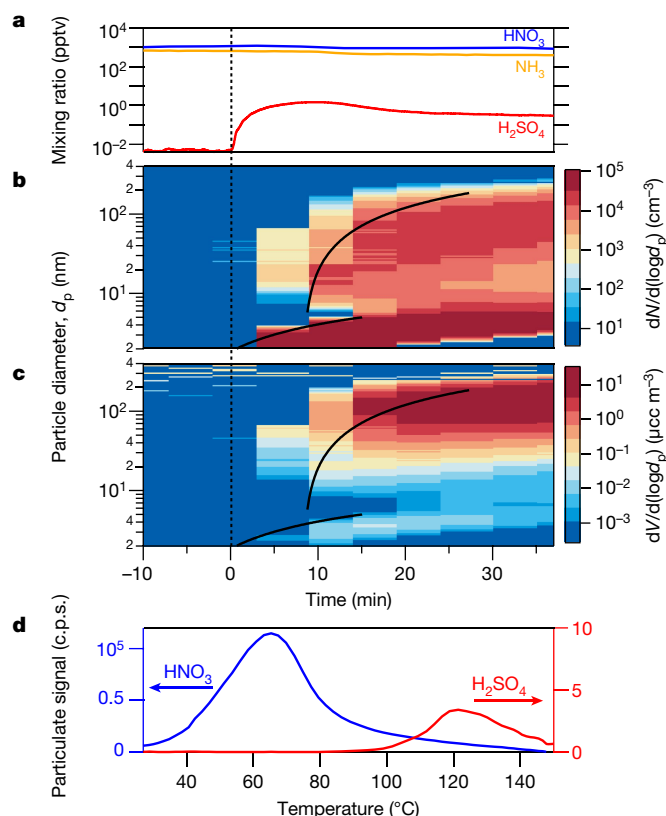


Fig. 2 | Chemical composition during a rapid growth event at +5 °C and 60% relative humidity. This growth event is indicated in Fig. 1c with a black-outlined purple square. **a**, Gas-phase nitric acid (NO_3), ammonia (NH_3) and sulfuric acid (H_2SO_4) mixing ratios versus time in an event initiated by SO_2 oxidation, with constant nitric acid and ammonia. **b**, Particle diameters and number distributions versus time, showing a clean chamber (to the left of the vertical dotted line), then nucleation after sulfuric acid formation and rapid growth once particles reach 2.3 nm. Black curves are linear fits to the 50% appearance times. **c**, Particle volume distributions from the same data, showing that 200-nm particles dominate the mass after 15 min. $1 \mu\text{cc} = 1 \text{ cm}^3$. **d**, FIGAERO thermogram from a 30-min filter sample after rapid growth (c.p.s., counts per second). The particle composition is dominated by nitrate with a core of sulfate, consistent with rapid growth by ammonium nitrate condensation on an ammonium sulfate (or bisulfate) core (note the different y-axis scales; the instrument is not sensitive to ammonia). A thermogram from just before the formation event shows no signal from either nitrate or sulfate, indicating that vapour adsorption did not interfere with the analysis.

‘nitric acid limited’, with more ammonia than nitric acid. All of these concentrations are well within the ranges typically observed in wintertime megacity conditions²⁴.

For both +5 °C and -10 °C, we consistently observe a relationship between S and d_{act} (we never achieved saturation at +20 °C and did not observe rapid growth). We observe no activation for S values of less than 1, and activation for S values greater than 1, with $\log d_{\text{act}}$ being inversely proportional to $\log([\text{HNO}_3] \cdot [\text{NH}_3])$ at each temperature (Extended Data Fig. 2b). Notably, d_{act} can be well under 10 nm and as low as 1.6 nm. This suggests that nitric acid and ammonia (ammonium nitrate) condensation may play a role in new-particle formation and growth within the valley of death, where very small particles are most vulnerable to loss by coagulation⁸.

We also performed experiments with only nitric acid, ammonia and water vapour added to the chamber (sulfuric acid contamination was measured to be less than 2×10^{-3} pptv). For temperatures of less than -15 °C and S values of more than 10^3 , we observed nucleation

and growth of pure ammonium nitrate particles (Fig. 3c). We progressively cooled the chamber to -24 °C, while holding the vapours at a constant level (Fig. 3b). The particle-formation rate ($J_{1.7}$) rose steadily from $0.006 \text{ cm}^{-3} \text{ s}^{-1}$ to $0.06 \text{ cm}^{-3} \text{ s}^{-1}$ at -24 °C. In Extended Data Fig. 3 we show a pure ammonium nitrate nucleation experiment performed at -25 °C under vapour conditions reported for the tropical upper troposphere⁴ (30–50 pptv nitric acid and 1.8 ppbv ammonia), showing that this mechanism can produce several 100 cm^{-3} particles per hour.

Our experiments show that semivolatile ammonium nitrate can condense on tiny nanoparticles, consistent with nano-Köhler theory²³. To confirm this we conducted a series of simulations using the monodisperse thermodynamic model MABNAG (model for acid-base chemistry in nanoparticle growth)²⁵, which treats known thermodynamics, including curvature (Kelvin) effects for a single evolving particle size. We show the points of the MABNAG simulations as triangles in Fig. 3a. MABNAG consistently and quantitatively confirms our experimental findings: there is little ammonium nitrate formation at S values of less than 1.0, as expected; and activation behaviour with ammonium nitrate condensation ultimately dominating the particle composition occurs at progressively smaller d_{act} values as S rises well above 1.0. The calculated and observed d_{act} values are broadly consistent. In Fig. 4 we show two representative MABNAG growth simulations for the two points indicated with open and filled diamonds in Fig. 3a; the simulations show no ammonium nitrate formation when conditions are undersaturated, but substantial formation when conditions are saturated, with activation behaviour near the observed $d_{\text{act}} = 4.7 \text{ nm}$. We show the calculated composition as well as diameter versus time for these and other cases in Extended Data Fig. 4.

We also conducted nano-Köhler simulations²³, shown in Extended Data Fig. 2c, which confirm the activation of ammonium nitrate condensation at diameters less than 4 nm, depending on the size of an assumed ammonium sulfate core. For a given core size the critical supersaturation required for activation at -10 °C is a factor of two to three higher than at +5 °C, consistent with the observed behaviour shown in Fig. 3a. While particles of 1–2 nm contain only a handful of acid and base molecules, the MABNAG and nano-Köhler simulations based on bulk thermodynamics—with only a Kelvin term to represent the unique behaviour of the nanoparticles—capture the activation and growth behaviours we observe.

Atmospheric implications

Our findings suggest that the condensation of nitric acid and ammonia onto nanoparticles to form ammonium nitrate (or, by extension, aminium nitrates in the presence of amines) may be important in the atmosphere. This process may contribute to urban new-particle formation during wintertime via rapid growth. It may also play a role in free-tropospheric particle formation, where sufficient vapours may exist to allow nucleation and growth of pure ammonium nitrate particles. We observe these behaviours in CLOUD for vapour concentrations well within those typical of the atmosphere.

Rapid growth may contribute to the often puzzling survival of newly formed particles in megacities, where particles form at rates consistent with sulfuric-acid-base nucleation and appear to grow at typical rates (roughly 10 nm h^{-1}) in the presence of extremely high condensation sinks that seemingly should scavenge all of the tiny nucleated particles. As shown in Extended Data Fig. 1, the ratio of $10^4 \times$ condensation sink (CS; in units of s^{-1}) to growth rate (GR; in nm h^{-1}) during nucleation events in Asian megacities typically ranges between 20 and 50, where the survival probability of particles with sizes of between 1.5 nm and 3 nm should drop precipitously³. However, the observed growth rates are based on appearance times in

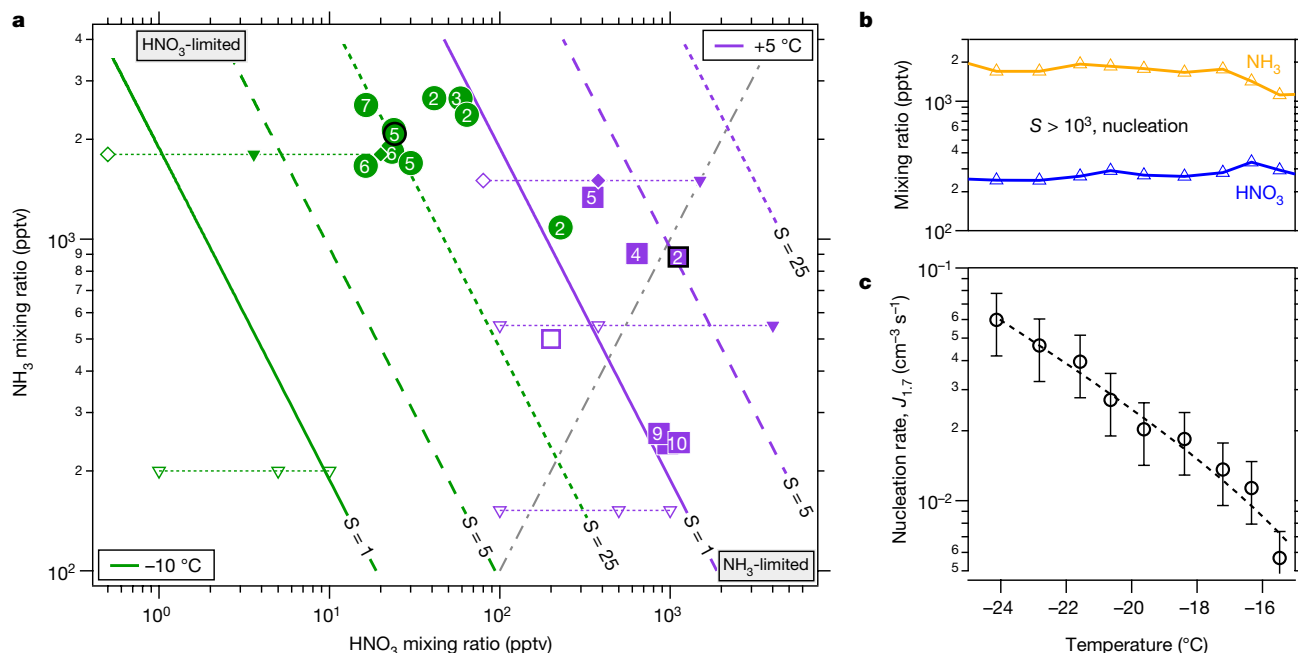


Fig. 3 | Phase space for rapid growth and nucleation. **a**, Ammonium nitrate saturation ratios versus gas-phase nitric acid and ammonia mixing ratios at 60% relative humidity. The coloured lines (slope = -1) represent $S=1$ (bold), $S=5$ (dashed) and $S=25$ (dotted), at $-10\text{ }^{\circ}\text{C}$ (green) and $+5\text{ }^{\circ}\text{C}$ (purple). The slope = +1 dot-dashed grey line indicates a 1:1 ammonia:nitric acid stoichiometry; the phase space to the upper left of this line is nitric-acid limited. Observed activation diameters (in nm) for measured nitric-acid–ammonia pairs are plotted as numbers inside solid circles and squares; open symbols show no activation. Activation occurs only for S values of more than 1, and the activation diameter decreases as S increases. Points from MABNAG simulations are shown with open triangles for no activation and filled triangles for activation;

measured ambient size distributions—just as in Figs. 1, 2—and thus reflect a spatial and temporal average of air masses passing over a sampling site during the course of a day. Rapid growth rates can reduce CS:GR by a factor of ten or more, effectively displacing urban ratios into a range characteristic of remote regions (Extended Data Fig. 1b). The empirically derived nucleation rates in Extended Data Fig. 1b correlate positively with high CS:GR, consistent with high production rates of condensable vapours; however, the complicated microphysics of particles smaller than 10 nm make a simple determination of the growth rate difficult. Urban conditions are however far less homogeneous than those of CLOUD, or even of remote boreal forests such as Hyytiälä. Because survival probability depends exponentially on CS:GR (refs. 3,7), but spatial (and temporal) averaging as well as ambient mixing are linear, real urban conditions may contain pockets conducive to transient rapid growth and thus unusually high survival probability that are blurred in the (averaged) observations.

The key here is that nitric acid vapour and ammonia are often at least one thousand times more abundant than sulfuric acid vapour. Thus, although they tend towards equilibrium with ammonium nitrate in the particle phase, even a modest perturbation above saturation can unleash a tremendous thermodynamic driving force for condensational growth, nominally up to one thousand times faster than growth by sulfuric acid condensation. This may be brief, but because of the disparity in concentrations, even a small deviation in saturation ratio above 1.0 may drive rapid growth for a short period at several nanometres per minute, as opposed to several nanometres per hour. The particles will not experience rapid growth for long, but they may grow fast enough to escape the valley of death.

simulations indicated with diamonds are shown in detail in Fig. 4 and Extended Data Fig. 4. Points from runs shown in Figs. 1, 2 are emphasized with a thick black outline. **b**, Mixing ratios for ammonia and nitric acid vapour during a pure ammonium nitrate nucleation scan from $-16\text{ }^{\circ}\text{C}$ to $-24\text{ }^{\circ}\text{C}$. **c**, Particle-formation (nucleation) rates ($J_{1,7}$) during the nucleation scan, showing a strong inverse relationship with temperature at constant HNO_3 and NH_3 , with H_2SO_4 concentrations of less than 0.002 pptv and relative humidity starting at 60% and ending at 40%. The bars indicate 30% estimated total errors on the nucleation rates, although the overall systematic scale uncertainties of $\pm 10\%$ on the NH_3 mixing ratio and $\pm 25\%$ on the HNO_3 mixing ratio are not shown.

We illustrate rapid growth in Fig. 4. Under most urban conditions, nucleation and early growth up to the activation size are likely to be controlled by sulfuric acid and a base (ammonia or an amine), shown by the red ‘cores’ in Fig. 4b. During the day (even in wintertime)—when NO_2 is oxidized by OH in the gas phase to produce nitric acid at rates of up to 3 ppbv h^{-1} , and ammonia from traffic, other combustion emissions and agriculture can reach 8 ppbv (ref. 24)—nitric acid and ammonia will not equilibrate, but rather will approach a modest steady-state supersaturation that drives ammonium nitrate formation to balance the production and emissions. However, this steady state will only be reached after several e-folding time periods set by the particle condensation sink. Typically, new-particle formation occurs at the lower end of the condensation-sink distribution (even under urban conditions)^{2,7}, so this timescale will be several minutes, or a length scale of hundreds of metres in the horizontal and tens of metres in the vertical. There are ample sources of inhomogeneity on this timescale, including inhomogeneous sources such as traffic on major roadways and vertical mixing (with an adiabatic lapse rate of $-9\text{ }^{\circ}\text{C km}^{-1}$)²⁴. Further, large eddy simulations of a megacity (Hong Kong) confirm widespread eddies with spatial scales of tens to hundreds of metres and velocity perturbations of the order 1 m s^{-1} (ref. 26). This is consistent with the sustained inhomogeneity required for the rapid growth we demonstrate here, shown conceptually in Fig. 4a. It is thus likely that dense urban conditions will typically include persistent inhomogeneities that maintain supersaturation of nitric acid and ammonia with sufficient magnitude to drive rapid growth, as indicated by the blue ‘shell’ in Fig. 4b. Our thermodynamic models support the phenomenology of Fig. 4b, as shown in Fig. 4c, d, although the composition is likely to be an amorphous mixture of salts (Extended Data Fig. 4).

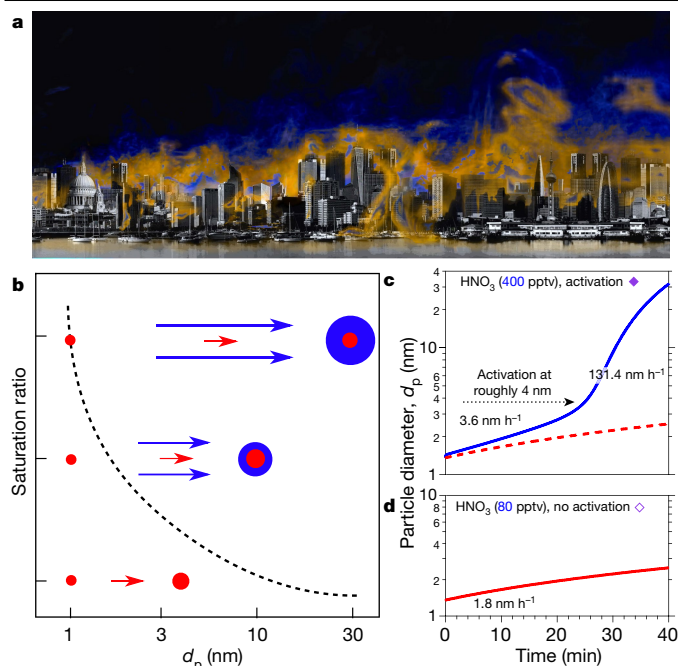


Fig. 4 | Conditions for rapid growth. Persistent supersaturations of ammonia and nitric acid with respect to ammonium nitrate will be sustained by inhomogeneity in urban conditions with high source strength. This will be sufficient to accelerate particle growth in the range 1–10 nm, where survival is threatened by the high coagulation sink of pre-existing particles from pollution. **a**, Conceptual image of urban conditions, where inhomogeneities in the concentrations of ammonia and nitric acid vapour and in temperatures are caused by non-uniform sources and large-scale eddies. **b**, Particles nucleate and grow slowly as (base-stabilized) sulfate (red). The activation size (shown with d_p on the x-axis) correlates inversely with the ammonium nitrate saturation ratio (shown qualitatively on the y-axis), as indicated by the dashed curve. Available concentrations of gas-phase nitric acid can exceed those of sulfuric acid by a factor of 1,000, so modest supersaturation drives rapid growth (blue) above an activation diameter determined by particle curvature (the Kelvin term). **c**, **d**, Monodisperse thermodynamic growth calculations (from MABNAG simulations) for high (**c**) and low (**d**) saturation ratios of ammonium nitrate, corresponding to **b** and to the closed and open diamonds towards the upper right in Fig. 3a. For a saturation ratio near 4, activation is predicted to occur near 4 nm, consistent with our observations.

Rapid growth may be sufficient for particles to grow from vulnerable sizes near 2.5 nm to more robust sizes larger than 10 nm. For example, repeated nucleation bursts with very rapid growth were observed in the ammonia- and nitric-acid-rich Cabauw site in the Netherlands during the EUCAARI campaign²⁷.

It is common for chemical transport models to use an equilibrium assumption for ammonium nitrate partitioning, because—on the time-scale of the coarse spatial grids and long time steps characteristic of large-scale models—the ammonium nitrate aerosol system should equilibrate with respect to the bulk submicrometre-size particles. Further, because rapid growth appears to be rate limited by the formation of ammonium nitrate, the covariance of base and nitric acid sources and concentrations may be essential. Even typical megacity steady-state vapour concentrations fall somewhat above the green points in Fig. 3a (towards larger mixing ratios). For constant production rates, as the temperature falls the ammonium nitrate saturation lines shown in Fig. 3a will sweep from the upper right towards the lower left, moving the system from rough equilibrium for typical urban production and emission rates when it is warmer than about +5 °C, to a sustained supersaturation when it is colder. Just as equilibrium organic condensation and partitioning results in underestimated

growth rates from organics in the boreal forest²⁸, equilibrium treatments of ammonium nitrate condensation will underestimate the role of nitric acid in nanoparticle growth, especially for inhomogeneous urban environments.

Although the pure ammonium nitrate nucleation rates in Fig. 3c are too slow to compete in urban new-particle formation, this mechanism may provide an important source of new particles in the relatively clean and cold upper free troposphere, where ammonia can be convected from the continental boundary layer²⁹ and abundant nitric acid is produced by electrical storms⁴. Theoretical studies have also suggested that nitric acid may serve as a chaperone to facilitate sulfuric-acid–ammonia nucleation³⁰. Larger (60–1,000 nm) particles consisting largely of ammonium nitrate, along with more than 1 ppbv of ammonia, have been observed by satellite in the upper troposphere during the Asian monsoon anticyclone⁴, and abundant 3–7-nm particles have been observed in situ in the tropical convective region at low temperature and condensation sink⁵. Although these particles are probably formed via nucleation, the mechanism is not yet known. However, our experiment under similar conditions (Extended Data Fig. 3) shows that it is plausible that pure ammonium nitrate nucleation and/or rapid growth by ammonium nitrate condensation contributes to these particles in the upper troposphere.

Our results indicate that the condensation of nitric acid and ammonia is likely to be an important new mechanism for particle formation and growth in the cold upper free troposphere, as supported by recent observations^{4,5}. Furthermore, this process could help to explain how newly formed particles survive scavenging losses in highly polluted urban environments³. As worldwide pollution controls continue to reduce SO₂ emissions sharply, the importance of NO_x and nitric acid for new-particle formation is likely to increase. In turn, controls on NO_x and ammonia emissions may become increasingly important, especially for the reduction of urban smog.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2270-4>.

- Stanier, C. O., Khlystov, A. Y. & Pandis, S. N. Nucleation events during the Pittsburgh Air Quality Study: description and relation to key meteorological, gas phase, and aerosol parameters. *Aerosol Sci. Technol.* **38**, 253–264 (2004).
- Yao, L. et al. Atmospheric new particle formation from sulfuric acid and amines in a Chinese megacity. *Science* **361**, 278–281 (2018).
- Kulmala, M., Kerminen, V.-M., Petäjä, T., Ding, A. J. & Wang, L. Atmospheric gas-to-particle conversion: why NPF events are observed in megacities? *Faraday Discuss.* **200**, 271–288 (2017).
- Höpfner, M. et al. Ammonium nitrate particles formed in upper troposphere from ground ammonia sources during Asian monsoons. *Nat. Geosci.* **12**, 608–612 (2019).
- Williamson, C. J. et al. A large source of cloud condensation nuclei from new particle formation in the tropics. *Nature* **574**, 399–403 (2019).
- Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2013: The Physical Science Basis* (Cambridge Univ. Press, 2013).
- McMurry, P. H. et al. A criterion for new particle formation in the sulfur-rich Atlanta atmosphere. *J. Geophys. Res.* **D 110**, D22S02 (2005).
- Kulmala, M. et al. Direct observations of atmospheric aerosol nucleation. *Science* **339**, 943–946 (2013).
- Gordon, H. et al. Causes and importance of new particle formation in the present-day and pre-industrial atmospheres. *J. Geophys. Res.* **D 122**, 8739–8760 (2017).
- Riipinen, I. et al. Contribution of organics to atmospheric nanoparticle growth. *Nat. Geosci.* **5**, 453–458 (2012).
- Pierce, J. R. & Adams, P. J. Efficiency of cloud condensation nuclei formation from ultrafine particles. *Atmos. Chem. Phys.* **7**, 1367–1379 (2007).
- Kuang, C., McMurry, P. H. & McCormick, A. V. Determination of cloud condensation nuclei production from measured new particle formation events. *Geophys. Res. Lett.* **36**, L09822 (2009).
- Apte, J. S., Brauer, M., Cohen, A. J., Ezzi, M. & Pope, C. A. Ambient PM_{2.5} reduces global and regional life expectancy. *Environ. Sci. Technol. Lett.* **5**, 546–551 (2018).
- Chen, G., Wang, W.-C. & Chen, J.-P. Circulation responses to regional aerosol climate forcing in summer over East Asia. *Clim. Dyn.* **51**, 3973–3984 (2018).

15. Kerminen, V.-M. & Kulmala, M. Analytical formulae connecting the “real” and the “apparent” nucleation rate and the nuclei number concentration for atmospheric nucleation events. *J. Aerosol Sci.* **33**, 609–622 (2002).
16. Takahama, S., Wittig, A. E., Vayenas, D. V., Davidson, C. I. & Pandis, S. N. Modeling the diurnal variation of nitrate during the Pittsburgh Air Quality Study. *J. Geophys. Res.* **D 109**, D16S06 (2004).
17. Xu, W. et al. Changes in aerosol chemistry from 2014 to 2016 in winter in Beijing: insights from high-resolution aerosol mass spectrometry. *J. Geophys. Res.* **D 124**, 1132–1147 (2019).
18. McMurry, P. H. Photochemical aerosol formation from SO₂: a theoretical analysis of smog chamber data. *J. Colloid Interface Sci.* **78**, 513–527 (1980).
19. Kirkby, J. et al. Ion-induced nucleation of pure biogenic particles. *Nature* **533**, 521–526 (2016).
20. Stolzenburg, D. et al. Rapid growth of organic aerosol nanoparticles over a wide tropospheric temperature range. *Proc. Natl Acad. Sci. USA* **115**, 9122–9127 (2018).
21. O’Dowd, C. D. et al. Marine aerosol formation from biogenic iodine emissions. *Nature* **417**, 632–636 (2002).
22. Kirkby, J. et al. Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature* **476**, 429–433 (2011).
23. Kontkanen, J., Olenius, T., Kulmala, M. & Riipinen, I. Exploring the potential of nano-köhler theory to describe the growth of atmospheric molecular clusters by organic vapors using cluster kinetics simulations. *Atmos. Chem. Phys.* **18**, 13733–13754 (2018).
24. Lu, K. et al. Fast photochemistry in wintertime haze: consequences for pollution mitigation strategies. *Environ. Sci. Technol.* **53**, 10676–10684 (2019).
25. Yli-Juuti, T. et al. Model for acid-base chemistry in nanoparticle growth (MABNAG). *Atmos. Chem. Phys.* **13**, 12507–12524 (2013).
26. Letzel, M. O. et al. LES case study on pedestrian level ventilation in two neighbourhoods in Hong Kong. *Meteorol. Z. (Berl.)* **21**, 575–589 (2012).
27. Manninen, H. E. et al. EUCAARI ion spectrometer measurements at 12 European sites – analysis of new particle formation events. *Atmos. Chem. Phys.* **10**, 7907–7927 (2010).
28. Pierce, J. R. et al. Quantification of the volatility of secondary organic compounds in ultrafine particles during nucleation events. *Atmos. Chem. Phys.* **11**, 9019–9036 (2011).
29. Ge, C., Zhu, C., Francisco, J. S., Zeng, X. C. & Wang, J. A molecular perspective for global modeling of upper atmospheric NH₃ from freezing clouds. *Proc. Natl Acad. Sci. USA* **115**, 6147–6152 (2018).
30. Liu, L. et al. The role of nitric acid in atmospheric new particle formation. *Phys. Chem. Chem. Phys.* **20**, 17406–17414 (2018).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Mingyi Wang^{1,2,30}, Weimeng Kong^{3,30}, Ruby Marten⁴, Xu-Cheng He⁵, Dexian Chen¹⁶, Joschka Pfeifer⁷, Arto Heitto⁸, Jenni Kontkanen⁵, Lubna Dada⁵, Andreas Kürten⁹, Taina Yli-Juuti⁶, Hanna E. Manninen⁷, Stavros Amanatidis³, António Amorim¹⁰, Rima Baalbaki⁵, Andrea Baccarini⁴, David M. Bell⁴, Barbara Bertozzi¹¹, Steffen Bräkling¹², Sophia Brilke¹³, Lucía Caudillo Murillo⁹, Randall Chiu¹⁴, Biwu Chu⁵, Louis-Philippe De Menezes⁷, Jonathan Duplissy^{5,15}, Henning Finkenzeller¹⁴, Loic Gonzalez Carracedo¹³, Manuel Granzin⁹, Roberto Guida⁷, Armin Hansel^{16,17}, Victoria Hofbauer^{1,2}, Jordan Krechmer¹⁸, Katrianne Lehtipalo^{5,19}, Houssni Lamkaddam⁴, Markus Lampimäki⁵, Chuan Ping Lee⁴, Vladimir Makhmutov²⁰, Guillaume Marie⁹, Serge Mathot⁷, Roy L. Mauldin^{1,2,21}, Bernhard Mentler¹⁶, Tatjana Müller⁹, Antti Onnela⁷, Eva Partoll¹⁶, Tuukka Petäjä⁵, Maxim Philippov²⁰, Veronika Pospisilova⁴, Ananth Ranjithkumar²², Matti Rissanen^{5,23}, Birte Rörup⁵, Wiebke Scholz^{16,17}, Jiali Shen⁵, Mario Simon⁹, Mikko Sipilä⁵, Gerhard Steiner^{16,24}, Dominik Stolzenburg^{5,13}, Yee Jun Tham⁵, António Tomé²⁵, Andrea C. Wagner^{8,14}, Dongyu S. Wang⁴, Yonghong Wang⁵, Stefan K. Weber⁷, Paul M. Winkler¹³, Peter J. Wlasits¹³, Yusheng Wu⁵, Mao Xiao⁴, Qing Ye^{1,2,26}, Marcel Zauner-Wieczorek⁹, Xueqin Zhou⁴, Rainer Volkamer¹⁴, Ilona Riipinen²⁷, Josef Dommen⁴, Joachim Curtius⁹, Urs Baltensperger⁴, Vladimír Kulmala^{5,15,28,29}, Douglas R. Worsnop^{5,18}, Jasper Kirkby^{7,9}, John H. Seinfeld⁹, Imad El-Haddad⁴, Richard C. Flagan³ & Neil M. Donahue^{1,6,2,26}✉

¹Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, PA, USA.

²Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA. ³Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA.

⁴Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, Villigen, Switzerland. ⁵Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, Helsinki, Finland.

⁶Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.

⁷CERN, the European Organization for Nuclear Research, Geneva, Switzerland. ⁸Department of Applied Physics, University of Eastern Finland, Kuopio, Finland. ⁹Institute for Atmospheric and Environmental Sciences, Goethe University Frankfurt, Frankfurt am Main, Germany.

¹⁰CENTRA and Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Lisbon, Portugal. ¹¹Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany. ¹²Tofwerk, Thun, Switzerland. ¹³Faculty of Physics, University of Vienna, Vienna, Austria. ¹⁴Department of Chemistry and CIRES, University of Colorado at Boulder, Boulder, CO, USA. ¹⁵Helsinki Institute of Physics, University of Helsinki, Helsinki, Finland.

¹⁶Institute for Ion Physics and Applied Physics, University of Innsbruck, Innsbruck, Austria.

¹⁷Ionicon Analytik, Innsbruck, Austria. ¹⁸Aerodyne Research, Billerica, MA, USA. ¹⁹Finnish Meteorological Institute, Helsinki, Finland. ²⁰P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Moscow, Russia. ²¹Department of Atmospheric and Oceanic Sciences, University of Colorado at Boulder, Boulder, CO, USA. ²²School of Earth and Environment, University of Leeds, Leeds, UK. ²³Aerosol Physics Laboratory, Physics Unit, Faculty of Engineering and Natural Sciences, Tampere University, Tampere, Finland. ²⁴Grimm Aerosol Technik Ainring, Ainring, Germany. ²⁵Institute Infante Dom Luiz, University of Beira Interior, Covilhã, Portugal. ²⁶Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA. ²⁷Department of Applied Environmental Science, University of Stockholm, Stockholm, Sweden. ²⁸Joint International Research Laboratory of Atmospheric and Earth System Sciences, Nanjing University, Nanjing, China. ²⁹Aerosol and Haze Laboratory, Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing, China. ³⁰These authors contributed equally: Mingyi Wang, Weimeng Kong. ✉e-mail: nmd@andrew.cmu.edu

The CLOUD facility

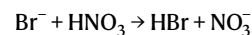
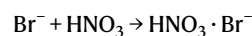
We conducted our measurements at the CERN CLOUD facility, a 26.1 m³ electropolished stainless-steel chamber that enables new-particle-formation experiments under the full range of tropospheric conditions with scrupulous cleanliness and minimal contamination^{22,31}. The CLOUD chamber is mounted in a thermal housing, capable of keeping temperature constant in a range between −65 °C and +100 °C with ±0.1 °C precision³², and relative humidity between 0.5% and 101%. Photochemical processes are initiated by homogeneous illumination with a built-in ultraviolet fibre-optic system, including four 200-W Hamamatsu Hg–Xe lamps at wavelengths between 250 nm and 450 nm and a 4-W KrF excimer ultraviolet laser at 248 nm with adjustable power. Ion-induced nucleation under different ionization levels is simulated with a combination of electric fields (±30 kV) and a high-flux beam of 3.6-GeV pions (π^+), which can artificially scavenge or enhance small ions. Uniform spatial mixing is achieved with magnetically coupled stainless-steel fans mounted at the top and bottom of the chamber. The characteristic gas mixing time in the chamber during experiments is a few minutes. The loss rate of condensable vapours and particles onto the chamber walls is comparable to the ambient condensation sink. To avoid contamination, the chamber is periodically cleaned by rinsing the walls with ultrapure water and heating to 100 °C for at least 24 h, ensuring extremely low contaminant levels of sulfuric acid (less than 5×10^4 cm^{−3}) and total organics (less than 50 pptv)^{19,33}. The CLOUD gas system is also built to the highest technical standards of cleanliness and performance. The dry air supply for the chamber is provided by boil-off oxygen (Messer, 99.999%) and boil-off nitrogen (Messer, 99.999%) mixed at the atmospheric ratio of 79:21. Highly pure water vapour, ozone and other trace gases can be precisely added at the pptv level.

Typical experimental sequence

To investigate the role of nitric acid in new-particle formation, we performed particle growth experiments at $T = -10$ °C, $+5$ °C and $+20$ °C, and (for the most part) at relative humidities of approximately 60%. A typical experiment started with illumination of the chamber at constant ozone (O₃) to photochemically produce •OH radicals. The subsequent oxidation of premixed SO₂, NO₂ and anthropogenic volatile organic compounds (VOCs; that is, toluene or cresol) led to the production of H₂SO₄, HNO₃ and highly oxygenated organic molecules (HOMs), respectively. As a result, nucleation was induced, followed (once the particles reached an activation diameter, d_{act}) by rapid growth via condensation of nitric acid and ammonia. In some experiments, we also injected nitric acid vapour directly into the CLOUD chamber from an ultrapure source to cover a wide range of conditions. In addition, to prove consistency we also carried out experiments with a biogenic precursor, α -pinene, replacing anthropogenic VOCs, as well as experiments without any organic vapours. For the experiments we focus on here, the HOM concentrations were either zero or small enough to have a minor effect on the experiment. In one case, we cooled the particle-free chamber (with fewer than 1 particle per cm^{−3}) continuously from −10 °C to −25 °C, while holding nitric acid and ammonia at a constant level, but with no sulfuric acid (less than 5×10^4 cm^{−3} or 2×10^{-3} pptv). We observed new-particle formation purely from nitric acid and ammonia at temperatures of −15 °C and lower. The nucleation rate grew as the temperature dropped. Moreover, as shown in Extended Data Fig. 3, at −25 °C new-particle formation events appeared to be quenched when we swept out primary ions with the electric field, and did not return until the field was turned off to allow primary ion production by galactic cosmic rays to again accumulate (roughly 700 cm^{−3}). We list the chamber conditions and key parameters for all the experiments here in Extended Data Table 1.

Instrumentation

To measure gas-phase nitric acid, we deployed a bromide chemical ionization atmospheric pressure interface time-of-flight (CI-API-TOF) mass spectrometer^{34,35} equipped with a commercial inlet (Airmodus) to minimize wall contact of the sample³⁶. We flowed dibromomethane (CH₂Br₂) into the ion-molecule reaction inlet to produce the primary reagent ion Br[−]. During its collision with HNO₃, Br[−] reacts either to form a cluster or via a proton transfer from the HNO₃ to form NO₃[−]:



To take the variation in the total reagent ions into account, we quantified nitric acid concentrations according to:

$$[\text{HNO}_3] = \frac{[\text{NO}_3^-]}{[\text{Br}^-] + [\text{H}_2\text{O} \cdot \text{Br}^-]} \times C$$

where C (in units of pptv) is a calibration coefficient obtained by measuring HNO₃/N₂ mixtures with known nitric acid concentrations. The nitric acid source was a portable permeation tube, kept constantly at 40 °C. An N₂ flow of 2 litres per minute was introduced into the permeation device to carry out the nitric acid vapour. To determine the permeation rate of nitric acid, we passed the outflow of the permeation tube through an impinger containing deionized water, and analysed the resulting nitric acid solution by spectrophotometry. Line losses during experiments and calibration procedures were calculated separately. We determined the corrected calibration coefficient to be 7,364 pptv.

Gas-phase ammonia was measured by a water cluster CI-API-TOF mass spectrometer (described elsewhere³⁷). The crossflow ion source coupled to a TOF mass spectrometer enables the selective measurement of basic compounds (for example, ammonia) by using positively charged water clusters as primary ions. Owing to the low reaction times (less than 1 ms), the instrument responds rapidly to changing chamber conditions with a detection limit of ammonia of 0.5 pptv.

Gas-phase sulfuric acid and HOMs were routinely measured with a detection limit of approximately 5×10^4 cm^{−3} by two nitrate CI-API-TOF mass spectrometers. One instrument was equipped with the Airmodus inlet and an X-ray generator as the ion source; the other deployed a home-made inlet and a corona discharge for ion generation³⁸. An electrostatic filter was installed in front of each instrument to remove ions and charged clusters formed in the chamber. Sulfuric acid and HOMs were quantified following calibration and loss correction procedures described previously^{19,22,39}.

VOCs were monitored by a proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS; Ionicon Analytik); this also provides information about the overall cleanliness regarding VOCs in the chamber. The technique has been extensively described previously⁴⁰. Direct calibration using diffusion sources allows us to determine VOC mixing ratios with an accuracy of 5% and a typical detection limit of 25 pptv (ref. ⁴¹).

Gas monitors were used to measure ozone (O₃; Thermo Environmental Instruments TEI 49C), sulfur dioxide (SO₂; Thermo Fisher Scientific 42i-TLE) and nitric oxide (NO; ECO Physics, CLD 780TR). Nitrogen dioxide (NO₂) was measured using a cavity-attenuated phase-shift NO₂ monitor (CAPS NO₂, Aerodyne Research) and a homemade cavity-enhanced differential optical absorption spectroscopy (CE-DOAS) instrument. The relative humidity of the chamber was determined using dew point mirrors (EdgeTech).

We measured the particle-phase composition via thermal desorption using an iodide-adduct chemical ionization time-of-flight mass spectrometer equipped with a filter inlet for gases and aerosols (FIGAERO-CIMS)^{42,43}. FIGAERO is a manifold inlet for a CIMS with

two operating modes. In one mode, gases are directly sampled into a 100-mbar turbulent flow ion-molecule reactor while particles are concurrently collected on a polytetrafluorethylene (PTFE) filter via a separate dedicated port. In the other mode, the filter is automatically moved into a pure N₂ gas stream flowing into the ion-molecule reactor while the N₂ is progressively heated to evaporate the particles via temperature-programmed desorption. Analytes are then chemically ionized and extracted into a TOF-MS, achieving a detection limit below 10⁶ cm⁻³.

Particle-size distributions between 1.8 nm and 500 nm were monitored continuously by a differential mobility analyser train (DMA-Train), a nano-scanning electrical mobility spectrometer (nSEMS), a nano-scanning mobility particle sizer (nano-SMPS), and a long-SMPS. The DMA-Train was constructed with six identical DMAs operating at different, but fixed, voltages. Particles transmitted through the DMAs were then detected by either a particle-size magnifier (PSM) or a CPC, depending on the size ranges. An approximation of the size distribution with 15 size bins was acquired by logarithmic interpolation between the six channels⁴⁴. The nSEMS used a new, radial opposed migration ion and aerosol classifier (ROMIAC), which is less sensitive to diffusional resolution degradation than the DMAs⁴⁵, and a soft X-ray charge conditioner. After leaving the classifier, particles were first activated in a fast-mixing diethylene glycol stage⁴⁶, and then counted with a butanol-CPC. The nSEMS transfer function that was used to invert the data to obtain the particle-size distribution was derived using three-dimensional finite-element modelling of the flows, electric field and particle trajectories^{47,48}. The two commercial mobility particle-size spectrometers, nano-SMPS and long-SMPS, have been fully characterized, calibrated and validated in several previous studies^{49–51}.

Determination of growth rate

The combined particle-size distribution was reconstructed using measurement data from DMA-Train at 1.8–4.3 nm, nSEMS at 4.3–18.1 nm, nano-SMPS at 18.1–55.2 nm and long-SMPS at 55.2–500 nm, and synchronized with long-SMPS measurement time. We list the sizing and resolution information of these instruments in Extended Data Table 2. As depicted in Extended Data Fig. 5a, the four instruments showed excellent agreement in their overlapping regions of the size ranges. The total number concentration obtained by integrating the combined size distribution agreed well with measurement by an Airmodus A11 nano-condensation nucleus counter (nCNC) and a TSI 3776 ultrafine condensation particle counter (UCPC) (Extended Data Fig. 5b). Particle growth rate, dd_p/dt , was then determined from the combined size distributions using the 50% appearance time method²⁰, as a clear front of a growing particle population could be identified during most rapid growth events (Extended Data Fig. 6). For the rapid growth rates, which are the principal focus here, the SMPS measurements provided the major constraint.

Determination of activation diameter

The activation diameter (d_{act}) was interpreted as the size at which growth accelerated from the slow, initial rate to the rapid, post-activation rate. The activation diameter was determined using the particle-size distribution acquired from DMA-Train or nSEMS at small sizes (less than 15 nm). At the activation diameter, the growth rate calculated from the 50% appearance time usually experienced a sharp change, from below 10 nm h⁻¹ to (often) over 100 nm h⁻¹, depending on concentrations of supersaturated HNO₃ and NH₃ vapours. A fast growth rate leads to a relatively low steady-state concentration of particles just above the activation diameter; the activation event often resulted in a notable gap in the particle-number size distribution. In some cases, a clear bimodal distribution was observed, with the number concentration in one size bin plunging below 10 counts cm⁻³, while the counts at larger sizes rose to more than 100 counts cm⁻³; the centroid diameter of the

size bin at which the number concentration dropped was then defined as the activation diameter (Extended Data Fig. 2a).

Calculation of saturation ratio

We model the ammonium nitrate formed in the particle phase as solid in our particle growth experiments, given that the relative humidity (roughly 60%) was less than the deliquescence relative humidity (DRH), given by⁵²:

$$\ln(\text{DRH}) = \frac{723.7}{T} + 1.6954$$

The dissociation constant, K_p , is defined as the product of the equilibrium partial pressures of HNO₃ and NH₃. K_p can be estimated by integrating the van't Hoff equation⁵³. The resulting equation for K_p in units of ppb² (assuming 1 atm of total pressure)⁵⁴ is:

$$\ln K_p = 118.87 - \frac{24,084}{T} - 6.025 \ln T$$

The saturation ratio, S , is thus calculated by dividing the product of measured mixing ratios of HNO₃ and NH₃ by the dissociation constant. The dissociation constant is quite sensitive to temperature changes, varying over more than two orders of magnitude for typical ambient conditions. Several degrees of temperature drop can lead to a much higher saturation ratio, shifting the equilibrium of the system towards the particle phase drastically. As illustrated in Extended Data Fig. 7, with an adiabatic lapse rate of −9 °C km⁻¹ during fast vertical mixing, upward transport of a few hundred metres alone is sufficient for a saturated nitric acid and ammonia air parcel to reach the saturation ratio capable of triggering rapid growth at a few nanometres.

Determination of nucleation rate

The nucleation rate, $J_{1.7}$, is determined here at a mobility diameter of 1.7 nm (a physical diameter of 1.4 nm) using particle size magnifier (PSM). At 1.7 nm, a particle is normally considered to be above its critical size and, therefore, thermodynamically stable. $J_{1.7}$ is calculated using the flux of the total concentration of particles growing past a specific diameter (here at 1.7 nm), as well as correction terms accounting for aerosol losses due to dilution in the chamber, wall losses and coagulation. Details can be found in our previous work⁵⁵.

The MABNAG model

To compare our measurements to thermodynamic predictions (including the Kelvin term for curved surfaces), we used the model for acid-base chemistry in nanoparticle growth (MABNAG)²⁵. MABNAG is a monodisperse particle population growth model that calculates the time evolution of particle composition and size on the basis of concentrations of condensing gases, relative humidity and ambient temperature, considering also the dissociation and protonation between acids and bases in the particle phase. In the model, water and bases are assumed always to be in equilibrium state between the gas and particle phases. Mass fluxes of acids to and from the particles are determined on the basis of their gas phase concentrations and their equilibrium vapour concentrations. In order to solve for the dissociation- and composition-dependent equilibrium concentrations, MABNAG couples a particle growth model to the extended aerosol inorganics model (E-AIM)^{56,57}. Here, we assumed particles in MABNAG to be liquid droplets at +5 °C and −10 °C, at 60% relative humidity. The simulation system consisted of four compounds: water, ammonia, sulfuric acid and nitric acid. The initial particle composition in each simulation was 40 sulfuric acid molecules and a corresponding amount of water and ammonia according to gas-particle equilibrium on the basis of their gas concentrations. With this setting, the initial diameter was approximately 2 nm. Particle density and surface tension were set to 1,500 kg m⁻³ and

Article

0.03 nm^{-1} , respectively. In Extended Data Fig. 4, we show that MAGNAG computations confirm that nitric acid and ammonia at the measured concentrations can activate small particles and cause rapid growth, and also confirm that the activation diameter depends on the ammonium nitrate saturation ratio, consistent with our measured diameter (diamonds in Fig. 3a).

Nano-Köhler theory

To prove consistency, we also calculated the equilibrium saturation ratios of ammonium nitrate above curved particle surfaces according to nano-Köhler theory²³. This theory describes the activation of nanometre-sized inorganic clusters to growth by vapour condensation, which is analogous to Köhler theory describing the activation of cloud condensation nuclei (CCN) to cloud droplets. Here, we assumed seed particles of ammonium sulfate, and performed calculations for three seed-particle diameters ($d_s = 1.4\text{ nm}$, 2.0 nm and 2.9 nm) at $+5^\circ\text{C}$ and -10°C , and at 60% relative humidity. The equilibrium vapour pressures of HNO_3 and NH_3 over the liquid phase, and the surface tension and density of the liquid phase, were obtained from an E-AIM^{56,57}. The equilibrium saturation ratios of ammonium nitrate were calculated as described in the Methods section ‘Calculation of saturation ratio’, also including the Kelvin term. The resulting Köhler curves, showing the equilibrium saturation ratio as a function of particle diameter, are presented in Extended Data Fig. 2c. The maxima of each curve corresponds to the activation diameter (d_{act}); saturation ratios of 10–50 lead to d_{act} values of 3–5 nm, consistent with our measurements in Fig. 3a. We summarize detailed results in Extended Data Table 1.

Ambient nucleation and growth

In Extended Data Table 3 we compile ambient observations of nucleation rates, growth rates and the ambient condensation sink. In most cases these are derived from evolving particle-size distributions. We summarize these observations in Extended Data Fig. 1.

Data availability

The full dataset shown in the figures and tables is publicly available⁵⁸. All data shown in the figures and tables and additional raw data are available upon request from the corresponding author. Source data for Figs. 1–4 and Extended Data Figs. 1–7 are provided with the paper.

Code availability

Codes for the MABNAG and nano-Köhler simulations and for conducting the analysis presented here can be obtained upon request from the corresponding author.

31. Duplissy, J. et al. Effect of ions on sulfuric acid-water binary particle formation: 2. Experimental data and comparison with QC-normalized classical nucleation theory. *J. Geophys. Res.* **D 121**, 1752–1775 (2016).
32. Dias, A. et al. Temperature uniformity in the CERN CLOUD chamber. *Aerosol Meas. Tech.* **10**, 5075–5088 (2017).
33. Schnitzhofer, R. et al. Characterisation of organic contaminants in the CLOUD chamber at CERN. *Aerosol Meas. Techn.* **7**, 2159–2168 (2014).
34. Jokinen, T. et al. Atmospheric sulphuric acid and neutral cluster measurements using CI-API-TOF. *Atmos. Chem. Phys.* **12**, 4117–4125 (2012).
35. Junninen, H. et al. A high-resolution mass spectrometer to measure atmospheric ion composition. *Atmos. Meas. Tech.* **3**, 1039–1053 (2010).
36. Eisele, F. & Tanner, D. Measurement of the gas phase concentration of H_2SO_4 and methane sulfonic acid and estimates of H_2SO_4 production and loss in the atmosphere. *J. Geophys. Res.* **D 98**, 9001–9010 (1993).
37. Pfeifer, J. et al. Measurement of the gas phase concentration of amines and iodine species using protonated water cluster chemical ionization mass spectrometry. *Atmos. Meas. Tech.* <https://doi.org/10.5194/amt-2019-215> (2019).
38. Kürten, A., Rondo, L., Ehrhart, S. & Curtius, J. Performance of a corona ion source for measurement of sulfuric acid by chemical ionization mass spectrometry. *Atmos. Meas. Tech.* **4**, 437–443 (2011).
39. Tröstl, J. et al. The role of low-volatility organic compounds in initial particle growth in the atmosphere. *Nature* **533**, 527–531 (2016).
40. Breitenlechner, M. et al. PTR3: an instrument for studying the lifecycle of reactive organic carbon in the atmosphere. *Anal. Chem.* **89**, 5824–5831 (2017).
41. Gautrois, M. & Koppmann, R. Diffusion technique for the production of gas standards for atmospheric measurements. *J. Chromatogr. A* **848**, 239–249 (1999).
42. Wang, M. et al. Reactions of atmospheric particulate stabilized Criegee intermediates lead to high molecular weight aerosol components. *Environ. Sci. Technol.* **50**, 5702–5710 (2016).
43. Lopez-Hilfiker, F. D. et al. A novel method for online analysis of gas and particle composition: description and evaluation of a Filter Inlet for Gases and AEROSols (FIGAERO). *Atmos. Meas. Tech.* **7**, 983–1001 (2014).
44. Stolzenburg, D., Steiner, G. & Winkler, P. M. A DMA-train for precision measurement of sub-10 nm aerosol dynamics. *Atmos. Meas. Tech.* **10**, 1639–1651 (2017).
45. Mui, W., Mai, H., Downard, A. J., Seinfeld, J. H. & Flagan, R. C. Design, simulation, and characterization of a radial opposed migration ion and aerosol classifier (ROMIAC). *Aerosol Sci. Technol.* **51**, 801–823 (2017).
46. Wimmer, D. et al. Performance of diethylene glycol-based particle counters in the sub-3 nm size range. *Atmos. Meas. Tech.* **6**, 1793–1804 (2013).
47. Mai, H. & Flagan, R. C. Scanning DMA data analysis I. Classification transfer function. *Aerosol Sci. Technol.* **52**, 1382–1399 (2018).
48. Mai, H., Kong, W., Seinfeld, J. H. & Flagan, R. C. Scanning DMA data analysis II. Integrated DMA-CPC instrument response and data inversion. *Aerosol Sci. Technol.* **52**, 1400–1414 (2018).
49. Jurányi, Z. et al. A 17 month climatology of the cloud condensation nuclei number concentration at the high alpine site Jungfraujoch. *J. Geophys. Res.* **D 116**, D10204 (2011).
50. Tröstl, J. et al. Fast and precise measurement in the sub-20 nm size range using a scanning mobility particle sizer. *J. Aerosol Sci.* **87**, 75–87 (2015).
51. Wiedensohler, A. et al. Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions. *Atmos. Meas. Tech.* **5**, 657–685 (2012).
52. Seinfeld, J. H. & Pandis, S. N. *Atmospheric Chemistry and Physics* 2nd edn (John Wiley & Sons, 2006).
53. Denbigh, K. G. & Denbigh, K. G. *The Principles of Chemical Equilibrium: With Applications in Chemistry and Chemical Engineering* (Cambridge Univ. Press, 1981).
54. Mozurkewich, M. The dissociation constant of ammonium nitrate and its dependence on temperature, relative humidity and particle size. *Atmos. Environ.* **A 27**, 261–270 (1993).
55. Lehtipalo, K. et al. Multi-component new particle formation from sulfuric acid, ammonia and biogenic vapors. *Sci. Adv.* **4**, eaau5363 (2018).
56. Clegg, S. L. & Seinfeld, J. H. Thermodynamic models of aqueous solutions containing inorganic electrolytes and dicarboxylic acids at 298.15 K. 1. The acids as nondissociating components. *J. Phys. Chem. A* **110**, 5692–5717 (2006).
57. Clegg, S. L. & Seinfeld, J. H. Thermodynamic models of aqueous solutions containing inorganic electrolytes and dicarboxylic acids at 298.15 K. 2. Systems including dissociation equilibria. *J. Phys. Chem. A* **110**, 5718–5734 (2006).
58. Wang, M. et al. Rapid growth of new atmospheric particles by nitric acid and ammonia condensation: data resources. <https://doi.org/10.5281/zenodo.3653377> (2020).
59. Xiao, S. et al. Strong atmospheric new particle formation in winter in urban Shanghai, China. *Atmos. Chem. Phys.* **15**, 1769–1781 (2015).
60. Iida, K., Stolzenburg, M. R., McMurry, P. H. & Smith, J. N. Estimating nanoparticle growth rates from size-dependent charged fractions: Analysis of new particle formation events in Mexico City. *J. Geophys. Res.* **D Atmospheres** **113**, D05207 (2008).
61. Mordas, G. et al. On operation of the ultra-fine water-based CPC TSI 3786 and comparison with other TSI models (TSI 3776, TSI 3772, TSI 3025, TSI 3010, TSI 3007). *Aerosol Sci. Technol.* **42**, 152–158 (2008).
62. Lehtipalo, K. et al. The effect of acid-base clustering and ions on the growth of atmospheric nano-particles. *Nat. Commun.* **7**, 11594 (2016).
63. Dal Maso, M. et al. Aerosol size distribution measurements at four Nordic field stations: identification, analysis and trajectory analysis of new particle formation bursts. *Tellus B Chem. Phys. Meteorol.* **59**, 350–361 (2007).
64. Dal Maso, M. et al. Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland. *Boreal Environ. Res.* **10**, 323–326 (2005).
65. Komppula, M. et al. Observations of new particle formation and size distributions at two different heights and surroundings in subarctic area in northern Finland. *J. Geophys. Res.* **D Atmospheres** **108** (D9), 4295 (2003).
66. Vehkamäki, H. et al. Atmospheric particle formation events at Värriö measurement station in Finnish Lapland 1998–2002. *Atmos. Chem. Phys.* **4**, 2015–2023 (2004).
67. Dal Maso, M. et al. Aerosol particle formation events at two Siberian stations inside the boreal forest. *Boreal Environ. Res.* **13**, 81–92 (2008).
68. Hussein, T. et al. Observation of regional new particle formation in the urban atmosphere. *Tellus B Chem. Phys. Meteorol.* **60**, 509–521 (2008).
69. Pikridas, M. et al. In situ formation and spatial variability of particle number concentration in a European megacity. *Atmos. Chem. Phys.* **15**, 10219–10237 (2015).
70. Hamed, A. et al. Nucleation and growth of new particles in Po Valley, Italy. *Atmos. Chem. Phys.* **7**, 355–376 (2007).
71. Hama, S. M., Cordell, R. L., Kos, G. P., Weijers, E. & Monks, P. S. Sub-micron particle number size distribution characteristics at two urban locations in Leicester. *Atmos. Res.* **194**, 1–16 (2017).
72. Gao, J., Chai, F., Wang, T., Wang, S. & Wang, W. Particle number size distribution and new particle formation: new characteristics during the special pollution control period in Beijing. *J. Environ. Sci.* **24**, 14–21 (2012).
73. Wang, Z. et al. Characteristics of regional new particle formation in urban and regional background environments in the North China Plain. *Atmos. Chem. Phys.* **13**, 12495–12506 (2013).
74. Yue, D. et al. Characteristics of aerosol size distributions and new particle formation in the summer in Beijing. *J. Geophys. Res.* **D Atmospheres** **114**, D00G12 (2009).
75. Zhang, Y. et al. Characterization of new particle and secondary aerosol formation during summertime in Beijing, China. *Tellus B Chem. Phys. Meteorol.* **63**, 382–394 (2011).
76. Man, H. et al. Comparison of daytime and nighttime new particle growth at the HKUST supersite in Hong Kong. *Environ. Sci. Technol.* **49**, 7170–7178 (2015).

77. An, J. et al. Characteristics of new particle formation events in Nanjing, China: effect of water-soluble ions. *Atmos. Environ.* **108**, 32–40 (2015).
78. Herrmann, E. et al. Aerosols and nucleation in eastern China: first insights from the new SORPES-NJU station. *Atmos. Chem. Phys.* **14**, 2169–2183 (2014).
79. Yu, H. et al. Nucleation and growth of sub-3 nm particles in the polluted urban atmosphere of a megacity in China. *Atmos. Chem. Phys.* **16**, 2641–2657 (2016).
80. Peng, J. et al. Submicron aerosols at thirteen diversified sites in China: size distribution, new particle formation and corresponding contribution to cloud condensation nuclei production. *Atmos. Chem. Phys.* **14**, 10249–10265 (2014).
81. Kanawade, V. et al. Infrequent occurrence of new particle formation at a semi-rural location, Gadanki, in tropical Southern India. *Atmos. Environ.* **94**, 264–273 (2014).
82. Mönkkönen, P. et al. Measurements in a highly polluted Asian mega city: observations of aerosol number size distribution, modal parameters and nucleation events. *Atmos. Chem. Phys.* **5**, 57–66 (2005).
83. Kuang, C. et al. An improved criterion for new particle formation in diverse atmospheric environments. *Atmos. Chem. Phys.* **10**, 8469–8480 (2010).
84. Iida, K. et al. Contribution of ion-induced nucleation to new particle formation: Methodology and its application to atmospheric observations in Boulder, Colorado. *J. Geophys. Res. D Atmospheres* **111**, D23201 (2006).

Acknowledgements We thank the European Organization for Nuclear Research (CERN) for supporting CLOUD with technical and financial resources and for providing a particle beam from the CERN Proton Synchrotron. This research has received funding from the US National Science Foundation (NSF; grant numbers AGS1602086, AGS1801329 and AGS-1801280); a NASA graduate fellowship (grant number NASA-NNX16AP36H); a Carnegie Mellon University Scott Institute Visiting Fellows grant; the Swiss National Science Foundation (grant numbers 200021_169090, 200020_172602 and 20FI20_172622); the European Community (EC) Seventh Framework Programme and the European Union (EU) H2020 programme (Marie Skłodowska Curie ITN CLOUD-TRAIN grant number 316662 and CLOUD-MOTION grant number 764991); a European Research Council (ERC) Advanced Grant (number ATM-GP 227463); an ERC Consolidator Grant (NANODYNAMITE 616075); an ERC Starting Grant (GASPARCON 714621), the Academy of Finland (grants 306853, 296628, 316114 and 299544); the Academy of Finland Center of Excellence programme (grant 307331); the German Federal Ministry of Education

and Research (CLOUD-12 number 01LK1222A and CLOUD-16 number 01LK1601A); the Knut and Alice Wallenberg Foundation Wallenberg Academy Fellow project AtmoRemove (grant number 2015.0162); the Austrian Science Fund (grant number P 27295-N20); the Portuguese Foundation for Science and Technology (grant number CERN/FIS-COM/0014/2017); and the Presidium of the Russian Academy of Sciences ('High energy physics and neutrino astrophysics' 2015). The FIGAERO-CIMS was supported by a Major Research Instrumentation (MRI) grant for the US NSF (AGS-1531284), and by the Wallace Research Foundation. We thank H. Cawley for producing Fig. 4a.

Author contributions M.W., R.M., J. Dommen, U.B., J. Kirkby, I.E.-H. and N.M.D. planned the experiments. M.W., W.K., R.M., X.-C.H., D.C., J.P., A.K., H.E.M., S.A., A.B., S. Bräkling, S. Brilke, L.C.M., B.C., L.-P.D.M., J. Duplissy, H.F., L.G.C., M.G., R.G., A. Hansel, V.H., J.K., K.L., H.L., C.P.L., V.M., G.M., S.M., B.M., T.M., A.O., E.P., T.P., M.P., V.P., M.R., B.R., W.S., J.S., M. Simon, M. Sipilä, G.S., D.S., Y.J.T., A.T., R.V., A.C.W., D.S.W., Y. Wang, S.K.W., P.M.W., P.J.W., Y. Wu, Q.Y., M.Z.-W., X.Z., J. Kirkby, I.E.-H. and R.C.F. prepared the CLOUD facility or measuring instruments. M.W., W.K., R.M., X.-C.H., D.C., J.P., L.D., H.E.M., S.A., A.A., R.B., A.B., D.M.B., B.B., S. Bräkling, S. Brilke, R.C., H.F., L.G.C., M.G., V.H., J.S., J. Duplissy, H.L., M.L., C.P.L., V.M., G.M., R.L.M., B.M., T.M., E.P., V.P., A.R., M.R., B.R., W.S., M. Simon, G.S., D.S., Y.J.T., A.T., A.C.W., D.S.W., Y. Wang, S.K.W., P.M.W., P.J.W., Y. Wu, M.X., M.Z.-W., X.Z., J. Kirkby and I.E.-H. collected the data. M.W., W.K., R.M., X.-C.H., D.C., J.P., A. Heitto, J. Kontkanen, L.D., A.K., T.Y.-J., H.E.M., S.A., L.G.C., J.S., W.S., M. Simon, D.S., D.S.W., S.K.W., P.M.W., I.E.-H., R.C.F. and N.M.D. analysed the data. M.W., W.K., R.M., X.-C.H., D.C., A. Heitto, J. Kontkanen, T.Y.-J., H.E.M., D.M.B., H.L., D.S., R.V., M.X., I.R., J. Dommen, J.C., U.B., M.K., D.R.W., J. Kirkby, J.H.S., I.E.-H., R.C.F. and N.M.D. contributed to the scientific discussion. M.W., W.K., R.M., X.-C.H., D.C., J.P., A. Heitto, J. Kontkanen, T.Y.-J., I.R., J. Dommen, U.B., M.K., D.R.W., J. Kirkby, J.H.S., I.E.-H., R.C.F. and N.M.D. wrote the manuscript.

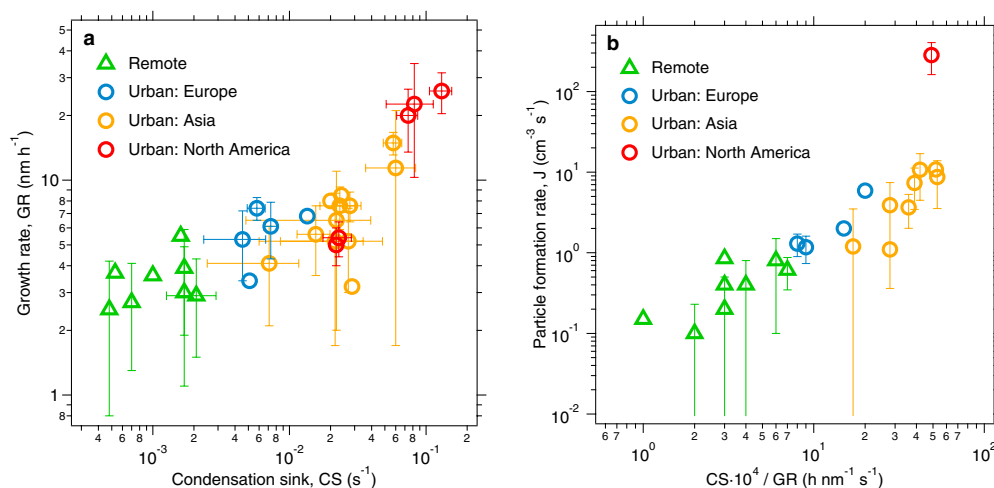
Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.D.

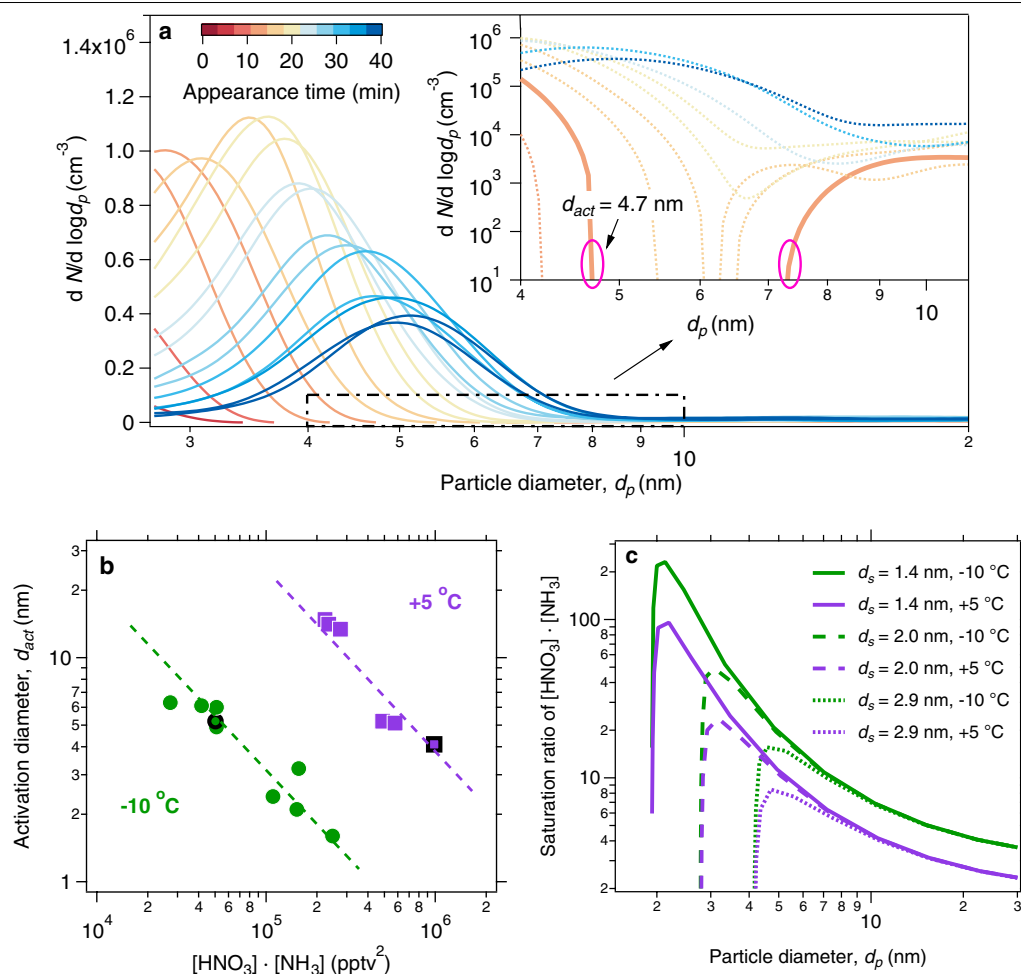
Peer review information *Nature* thanks Hugh Coe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | New-particle-formation events observed in various remote and urban environments (see Extended Data Table 3 for a complete set of references). a, Growth rates (GR) versus condensation sinks (CS), showing that both are higher in polluted urban environments than in other environments. **b,** Particle-formation rates (J) versus a measure of particle loss via coagulation ($CS \times 10^4 / GR$, similar to the the McMurry L parameter), showing

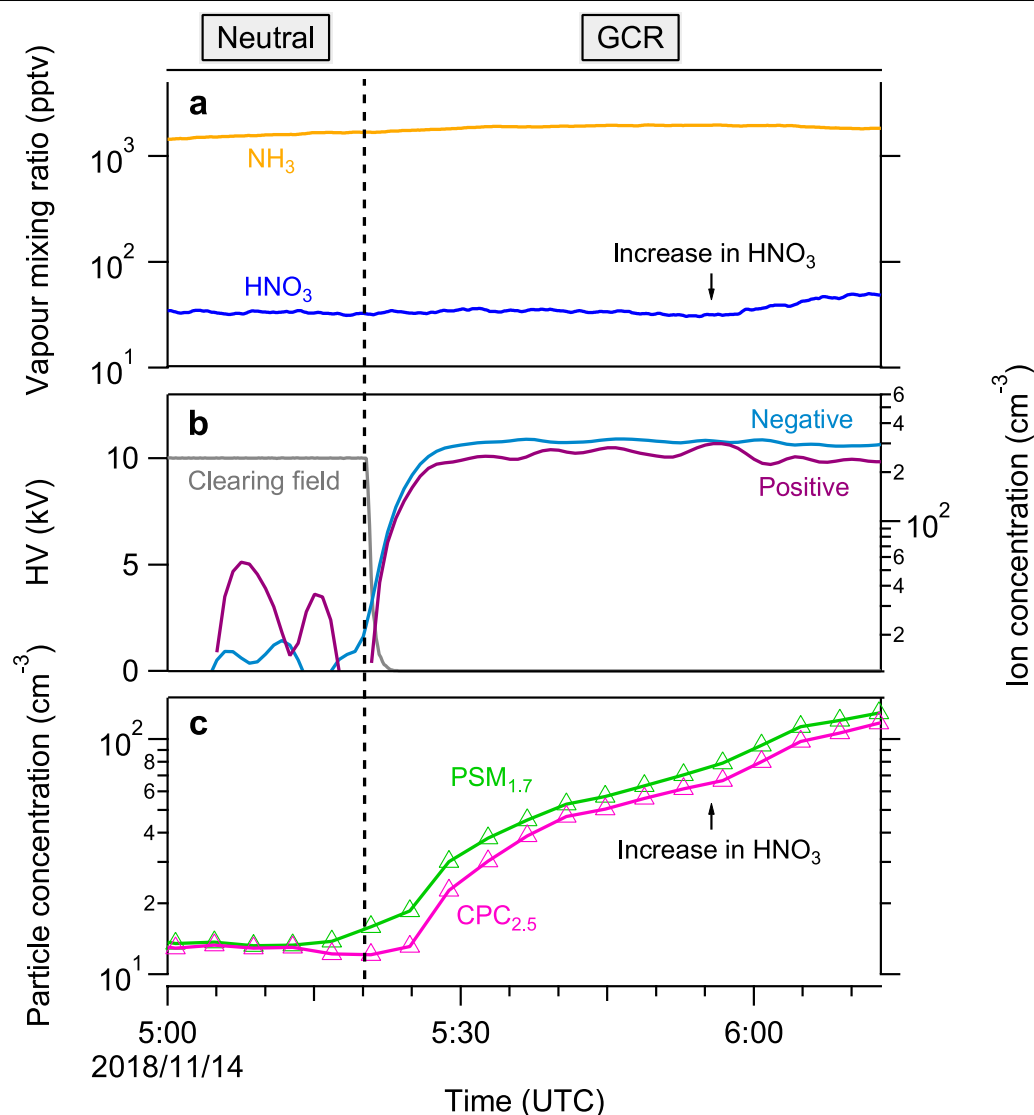
high new-particle-formation rates in urban conditions where the condensation sinks were so high compared to the growth rate that survival of nucleated particles should be very low. J and GR were calculated over the size range from a few nanometres to over 20 nm, except for J at Shanghai⁵⁹ and Tecamac⁶⁰, which were calculated from 3 nm to 6 nm. The bars indicate 1σ total errors.



Extended Data Fig. 2 | Activation diameter of newly formed particles.

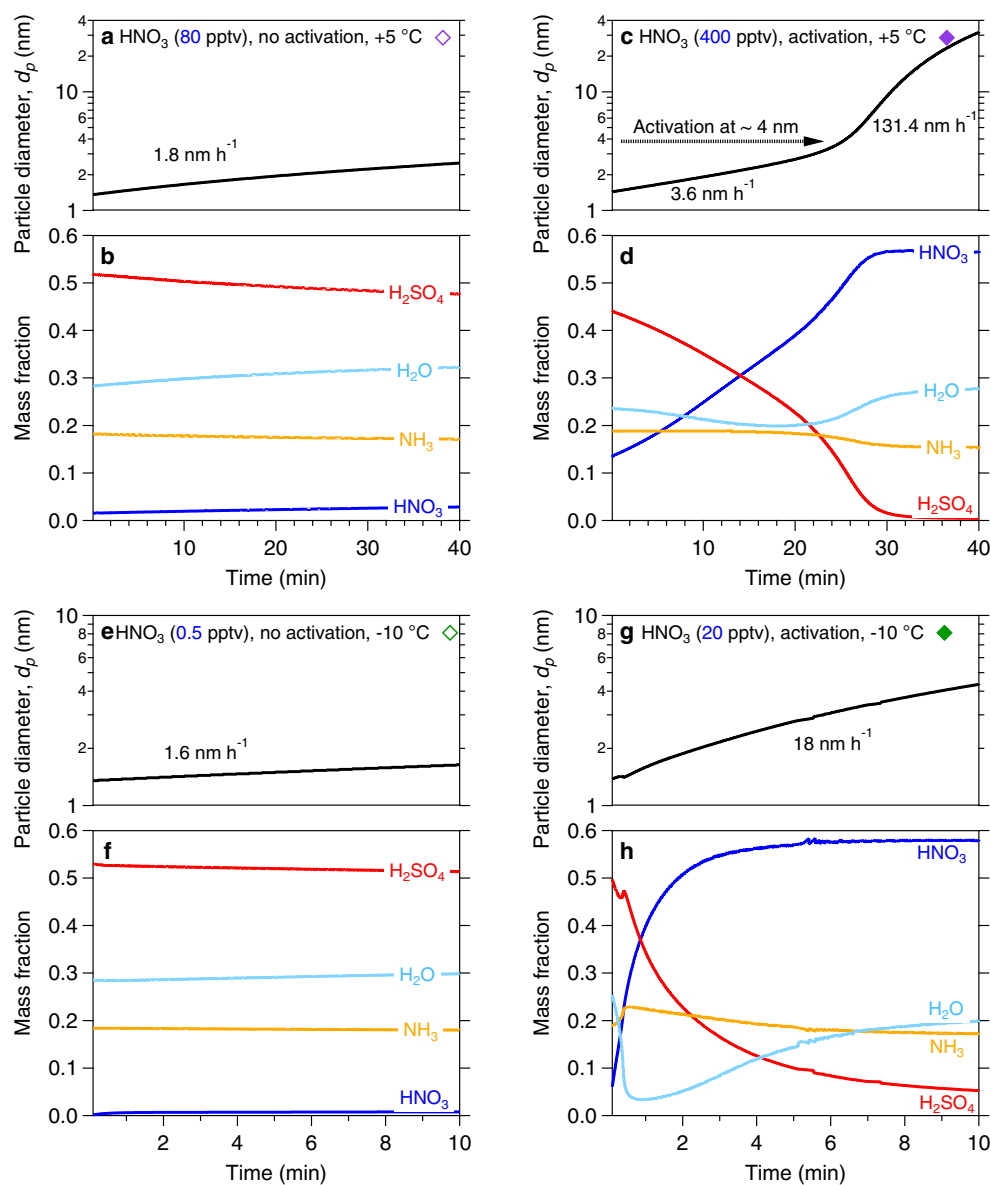
a, Determination of the activation diameter, d_{act} , from a rapid growth event at +5 °C, in the presence of nitric acid, ammonia and sulfuric acid. The solid orange trace in the insert indicates the first size distribution curve that exhibited a clear bimodal distribution, which appeared roughly 7 min after nucleation. We define the activation diameter as the largest observed size of the smaller mode. In this case, $d_{act} = 4.7$ nm, which agrees well with the value obtained from MABNAG simulations (roughly 4 nm) under the same conditions as in Fig. 4. **b**, Activation diameter versus vapour product. Measured activation diameters at a given temperature correlate inversely with the product of nitric

acid and ammonia vapours, in a log-log space. An amount of vapour product that is approximately one order of magnitude higher is required for the same d_{act} at +5 °C than at -10 °C, because of the higher vapour pressure (faster dissociation) of ammonium nitrate when it is warmer. **c**, Equilibrium particle diameter (d_p) at different saturation ratios of ammonium nitrate, calculated according to nano-Köhler theory. Purple curves are for +5 °C and green curves are for -10 °C (as throughout this work). The line type shows the diameter of the seed particle (d_s). The maximum of each curve corresponds to the activation diameter (d_{act}). A higher supersaturation is required for activation at lower temperature.



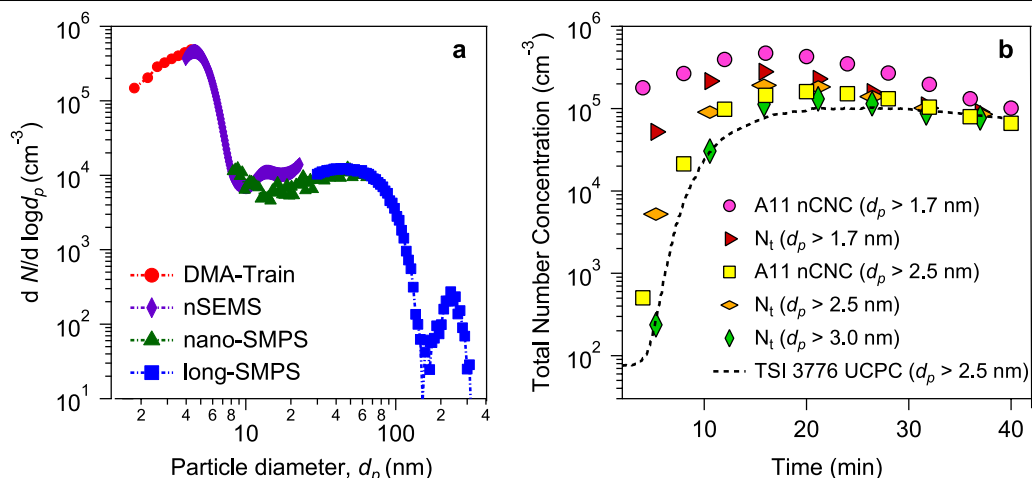
Extended Data Fig. 3 | A typical measurement sequence. Nucleation was carried out purely from nitric acid and ammonia, with no sulfuric acid (measured to less than $5 \times 10^4 \text{ cm}^{-3}$ or 2×10^{-3} pptv), as a function of coordinated universal time (UTC), at 60% relative humidity and -25°C . **a**, Gas-phase ammonia and nitric acid mixing ratios. The run started with injection of the nitric acid and ammonia flow into the chamber to reach chosen steady-state values near 30 pptv and 1,500 pptv, respectively. The nitric acid flow was increased at 5:53 on 14 November 2018 to prove consistency. **b**, Clearing-field voltage and ion concentrations. Primary ions were formed from galactic cosmic rays (GCR). The clearing-field high voltage (HV) was used to sweep out

small ions at the beginning of the run, and turned off at 05:21 on 14 November 2018 to allow the ion concentration to build up to a steady state between GCR production and wall deposition. **c**, Particle concentrations at two cut-off sizes (1.7 nm and 2.5 nm). Particles formed slowly in the chamber under 'neutral' conditions with the HV clearing field on and thus without ions present. The presence of ions ('GCR' condition) caused a sharp increase in the particle-number concentration by about one order of magnitude, with a slower approach to steady state because of the longer wall-deposition time constant for the larger particles. Particle numbers rose again with rising nitric acid.



Extended Data Fig. 4 | Comparison of growth rates and chemical composition in four simulations at +5 °C and -10 °C with the thermodynamic model MABNAG. The simulation points are shown in Fig. 3a (filled diamonds, with activation; open diamonds, without activation). **a, c, e, g.** Temporal evolution of the particle diameter. **b, d, f, h.** Temporal evolution of the particle-phase chemical composition. The left-hand column (**a, b, e, f**) shows simulations without activation. The right-hand column (**c, d, g, h**) shows simulations with activation. We set the HNO_3 mixing ratios at 80 pptv and 400 pptv with 1,500 pptv NH_3 at +5 °C, and set the HNO_3 mixing ratios at 20 pptv and 0.5 pptv with 1,500 pptv NH_3 at -10 °C, to simulate unsaturated

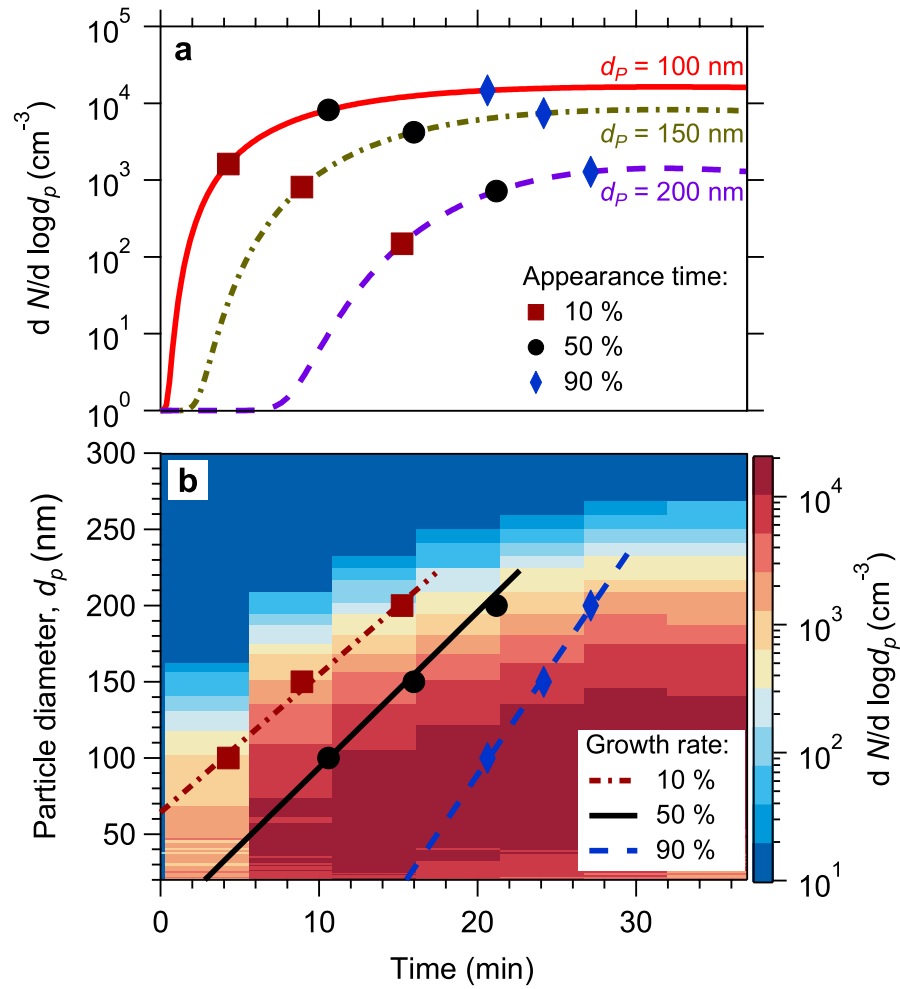
(**a, b, e, f**) and supersaturated (**c, d, g, h**) conditions, respectively. All other conditions were held constant for the simulations, with the $[\text{H}_2\text{SO}_4]$ at $2 \times 10^7 \text{ cm}^{-3}$ and relative humidity at 60%. Activation corresponds to a rapid increase in the nitric acid (nitrate) mass fraction; the simulations for activation conditions suggest that water activity may be an interesting variable influencing activation behaviour. The activated model results (**c, d, g, h**) confirm that supersaturated nitric acid and ammonia lead to rapid growth of nanoparticles. The simulated activation diameter at +5 °C is roughly 4 nm, similar to that from the chamber experiment (4.7 nm, Fig. 3a); at -10 °C the simulated activation diameter is less than 2 nm, smaller than observed.



Extended Data Fig. 5 | Combined particle-size distribution and total concentrations from four particle characterization instruments.

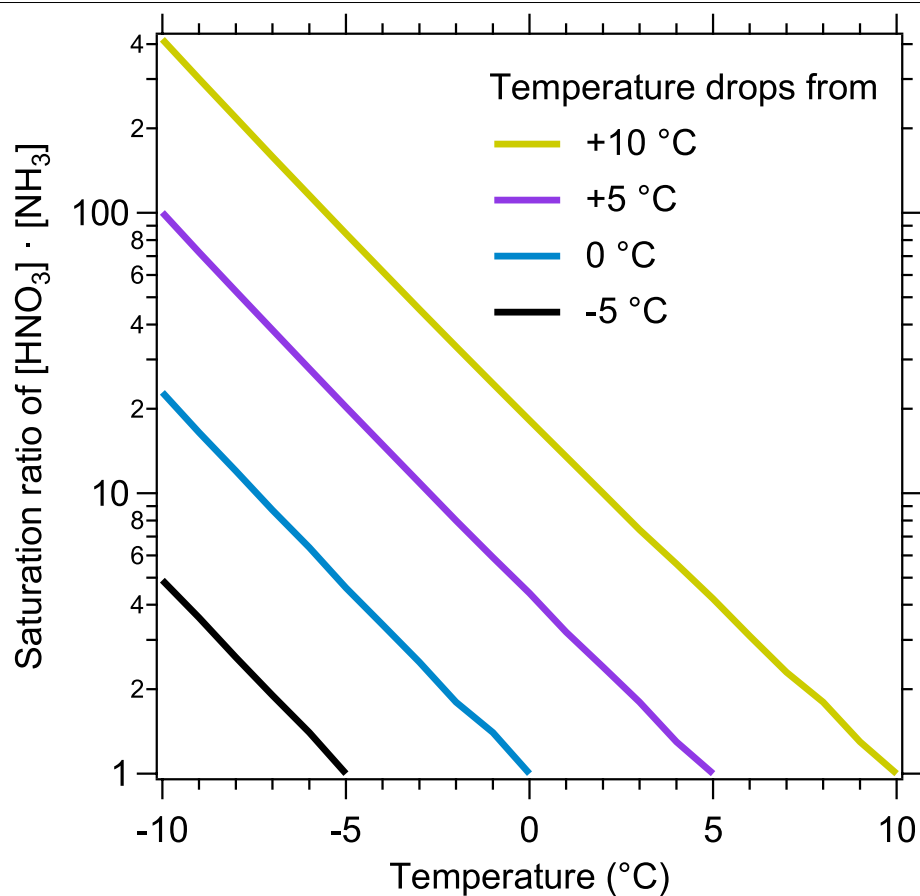
a, Combined size distributions, $n_N^o(d_p) = dN/d(\log d_p)$, from four electrical mobility particle-size spectrometers of different, but overlapping, detection ranges. The DMA-Train, nSEMS and nano-SMPS data were averaged every five minutes to coordinate with the long-SMPS scanning time resolution. The tail of the size distribution of large particles outside the detection range was extrapolated by fitting a lognormal distribution. **b**, Comparison of the integrated number concentrations from the combined size distributions in **a** with total number counts obtained from fixed-cutoff-size condensation particle counters. We obtained the total number concentration of particles, $N_t(d_{p0})$, above a cutoff size, d_{p0} , by integrating the particle-size distribution

using⁶²: $N_t = \int_{d_{p0}}^{\infty} \{n_N(d_p) \times \eta_{UCPC}\} dd_p$, applying the size-dependent detection efficiency, η_{UCPC} ⁶¹, to adjust the integrated total number concentration. We plot the total number concentrations for three different cutoff sizes (d_{p0}) of 1.7 nm, 2.5 nm and 3.0 nm, obtained every 5 min, with coloured symbols as shown in the legend. We also plot measured total number concentrations from two instruments: the Airmodus A11 nCNC-system at nominal cutoff sizes of 1.7 nm and 2.5 nm, and a TSI 3776 UCPC with a nominal cutoff size of 2.5 nm. The Airmodus A11 nCNC-system consists of an A10 PSM and an A20 CPC, which determined both the size distribution of 1–4-nm aerosol particles and the total number concentration of particles smaller than 1 μ m (ref. ⁶²). The TSI 3776 UCPC has a rapid response time and so, rather than the 5-min basis for the other points, we plot the values from this instrument with a dashed curve.



Extended Data Fig. 6 | Determination of growth rate using the appearance-time method. a, Logarithmic interpolated time-dependent growth profiles for particles with diameters of 100 nm, 150 nm and 200 nm. Three appearance times, when particle number concentrations reached 10%, 50%, and 90% of their maximum, are labelled with different symbols for the

three different diameters. **b,** Growth-rate calculation for a rapid growth event (as in Fig. 2) above the activation diameter. The growth rates, in nm h^{-1} , that we report here are the slopes of linear fits to the 50% (among 10%, 50% and 90%) appearance times calculated from all sizes above the activation diameter (the slope of the solid black line and the black circles in **b**).

**Extended Data Fig. 7 | Saturation ratio as a function of temperature.**

At constant nitric acid and ammonia, a decline in temperature leads to an exponential increase in the saturation ratio of ammonium nitrate, shown as the product of nitric acid and ammonia vapour concentration. With an adiabatic

lapse rate of -9 °C km^{-1} during adiabatic vertical mixing, upward transport of a few hundred metres alone is enough for a saturated nitric acid and ammonia air parcel to reach the saturation ratio capable of triggering rapid growth at a few nanometres.

Extended Data Table 1 | Conditions for all nucleation and growth experiments and nano-Köhler simulations discussed here

Run	T (°C)	HNO ₃ (pptv)	NH ₃ (pptv)	H ₂ SO ₄ (pptv)	HOMs (pptv)	RH %	Saturation ratio of [HNO ₃]·[NH ₃]	<i>d</i> _{act} (nm)
2163.01	+20	150	2400	3.60	n/a	60	0.03	n/a
2140.06	+5	860	259	0.96	5.18	60	1.18	9.4
2140.08	+5	971	242	0.84	5.29	60	1.25	10.4
2140.10	+5	1130	244	0.88	5.25	60	1.46	10.0
2170.05	+5	292	370	3.00	n/a	60	0.57	n/a
2170.15	+5	1117	883	1.48	n/a	60	5.24	2.3
2170.20	+5	352	1387	1.48	n/a	60	2.59	4.7
2174.06	+5	640 ^a	904	0.46	12.01	60	3.07	3.9
2156.13	-10	228	1085	0.16	n/a	60	131.44	1.6
2156.15	-10	64	2371	0.06	n/a	60	80.90	2.1
2156.17	-10	41	2659	0.07	n/a	60	58.31	2.4
2157.06	-10	n/a	1915	0.44	0.32	60	n/a	n/a
2158.02	-10	59 ^a	2654	0.38	0.26	60	82.97	2.9
2158.04	-10	24 ^a	2131	0.46	0.28	60	27.06	4.6
2158.06	-10	24 ^a	2077	0.36	0.29	60	26.72	4.5
2159.02	-10	23 ^a	1835	0.72	0.67	60	22.15	4.5
2159.04	-10	30 ^a	1694	1.04	0.78	60	27.07	4.9
2160.02	-10	16 ^a	1663	0.60	0.25	60	14.44	5.9
2160.06	-10	16 ^a	2535	0.30	0.23	60	22.11	6.5
2162.01	-15 → -24	272	1647	< 0.002	n/a	60 → 40	1252.22	Nucleation
<i>d</i> _s = 1.4 nm	+5	n/a	n/a	n/a	n/a	60	96	2.2
<i>d</i> _s = 2.0 nm	+5	n/a	n/a	n/a	n/a	60	24	3.2
<i>d</i> _s = 2.9 nm	+5	n/a	n/a	n/a	n/a	60	8	4.7
<i>d</i> _s = 1.4 nm	-10	n/a	n/a	n/a	n/a	60	231	2.1
<i>d</i> _s = 2.0 nm	-10	n/a	n/a	n/a	n/a	60	49	3.1
<i>d</i> _s = 2.9 nm	-10	n/a	n/a	n/a	n/a	60	16	4.6

^aHNO₃ production via NO₂ photo-oxidation.

n/a, not applicable; RH, relative humidity.

Extended Data Table 2 | Specifications of the four particle-sizing instruments used here

Instrument	Components	Size Range	Size Resolution	Time Resolution
DMA-Train	TSI 3776 UCPC TSI 3776 nanoEnhancer Airmodus A10 PSM	1.8 nm - 8.0 nm	15 bins (interpolated)	5 s
nSEMS	ROMIAC TSI 3760A CPC	1.5 nm - 23 nm	240 bins	1 min
nano-SMPS	TSI 3938 SMPS TSI 3776 UCPC	2 nm - 64 nm	96 bins	1 min
long-SMPS	TSI 3071 DMA TSI 3010 CPC	20 nm - 500 nm	84 bins	5 min

Extended Data Table 3 | Ambient particle-formation rates (J), growth rates (GR) and condensation sinks (CS) in various remote and urban environments

City/Region (Country)	J^a ($\text{cm}^{-3} \cdot \text{s}^{-1}$)	GR ^a ($\text{nm} \cdot \text{h}^{-1}$)	CS (s^{-1})	$\text{CS} \cdot 10^4 / \text{GR}$ ($\text{h} \cdot \text{nm}^{-1} \cdot \text{s}^{-1}$)	Ref.
Hyytiälä (Finland)	0.8 ± 0.7	3.0 ± 1.9	1.7×10^{-3}	6	63
Hyytiälä (Finland)	0.6 ± 0.3	2.9 ± 1.4	$2.1 \times 10^{-3} \pm 8.2 \times 10^{-4}$	7	64
Pallas (Finland)	0.1 ± 0.1	2.5 ± 1.7	4.8×10^{-4}	2	63
Pallas (Finland)	0.2	3.7	5.3×10^{-4}	1	65
Värriö (Finland)	0.9	3.6	1.0×10^{-3}	3	66
Värriö (Finland)	0.2 ± 0.3	2.7 ± 1.4	7.0×10^{-4}	3	63
Tomsk (Russia)	0.4	5.5	1.6×10^{-3}	3	67
Sörmland (Sweden)	0.4 ± 0.4	3.9 ± 2.0	1.7×10^{-3}	4	63
Helsinki (Finland)	2.0	3.4	5.1×10^{-3}	15	68
Paris (France)	n/a	6.1 ± 1.8	$7.3 \times 10^{-3} \pm 8.0 \times 10^{-4}$	12	69
Po Valley (Italy)	5.9	6.8	1.4×10^{-2}	20	70
Brookfield (UK)	1.2 ± 0.4	5.3 ± 1.9	$4.5 \times 10^{-3} \pm 2.2 \times 10^{-3}$	9	71
Leicester (UK)	1.3 ± 0.4	7.4 ± 0.9	$5.8 \times 10^{-3} \pm 8.7 \times 10^{-4}$	8	71
Beijing (China)	n/a	3.2	2.9×10^{-2}	90	72
Beijing (China)	10.8	5.2 ± 2.2	$2.7 \times 10^{-2} \pm 2.1 \times 10^{-2}$	52	73
Beijing (China)	10.7 ± 6.2	5.2 ± 3.5	$2.2 \times 10^{-2} \pm 1.3 \times 10^{-2}$	42	74
Beijing (China)	n/a	6.5 ± 4.5	$2.2 \times 10^{-2} \pm 1.7 \times 10^{-2}$	34	75
Hong Kong (China)	3.9 ± 3.5	5.6 ± 2.0	$1.6 \times 10^{-2} \pm 4.2 \times 10^{-3}$	28	76
Nanjing (China)	3.7 ± 1.6	7.6 ± 1.2	$2.8 \times 10^{-2} \pm 5.7 \times 10^{-3}$	36	77
Nanjing (China)	1.1	8.5	2.4×10^{-2}	28	78
Nanjing (China)	n/a	7.6 ± 1.7	$2.3 \times 10^{-2} \pm 6.7 \times 10^{-3}$	31	79
Shanghai (China)	8.7 ± 5.2^b	11.4 ± 9.7	$6.0 \times 10^{-2} \pm 2.4 \times 10^{-2}$	53	59
Shanghai (China)	n/a	8.0	2.0×10^{-2}	25	80
Gadanki (India)	1.2 ± 2.3	4.1 ± 2.0	$7.1 \times 10^{-3} \pm 4.6 \times 10^{-3}$	17	81
New Delhi (India)	7.3 ± 3.9	14.9 ± 1.8	$5.8 \times 10^{-2} \pm 8.9 \times 10^{-3}$	39	82
Tecamac (Mexico)	283.0 ± 121.0^b	26.0 ± 5.6	$1.3 \times 10^{-1} \pm 2.4 \times 10^{-2}$	49	60
Tecamac (Mexico)	n/a	22.6 ± 12.3	$8.2 \times 10^{-2} \pm 3.1 \times 10^{-2}$	36	60
Atlanta (US)	n/a	20.0 ± 6.5	$7.4 \times 10^{-2} \pm 1.3 \times 10^{-2}$	37	7, 83
Boulder (US)	n/a	5.4 ± 1.0	$2.3 \times 10^{-2} \pm 5.4 \times 10^{-3}$	42	84
Boulder (US)	n/a	5.0 ± 1.0	2.2×10^{-2}	43	60

^aJ and GR were mostly calculated over a size range from a few nanometres to more than 20 nm.

^bJ calculated from 3 nm to 6 nm.

Uncertainties indicate 1 σ errors. From refs. ^{759,60,63-84}.

Early Holocene crop cultivation and landscape modification in Amazonia

<https://doi.org/10.1038/s41586-020-2162-7>

Received: 21 November 2019

Accepted: 13 February 2020

Published online: 8 April 2020

 Check for updates

Umberto Lombardo^{1✉}, José Iriarte², Lautaro Hilbert³, Javier Ruiz-Pérez⁴, José M. Capriles^{5,6} & Heinz Veit¹

The onset of plant cultivation is one of the most important cultural transitions in human history^{1–4}. Southwestern Amazonia has previously been proposed as an early centre of plant domestication, on the basis of molecular markers that show genetic similarities between domesticated plants and wild relatives^{4–6}. However, the nature of the early human occupation of southwestern Amazonia, and the history of plant cultivation in this region, are poorly understood. Here we document the cultivation of squash (*Cucurbita* sp.) at about 10,250 calibrated years before present (cal. yr BP), manioc (*Manihot* sp.) at about 10,350 cal. yr BP and maize (*Zea mays*) at about 6,850 cal. yr BP, in the Llanos de Moxos (Bolivia). We show that, starting at around 10,850 cal. yr BP, inhabitants of this region began to create a landscape that ultimately comprised approximately 4,700 artificial forest islands within a treeless, seasonally flooded savannah. Our results confirm that the Llanos de Moxos is a hotspot for early plant cultivation and demonstrate that—ever since their arrival in Amazonia—humans have markedly altered the landscape, with lasting repercussions for habitat heterogeneity and species conservation.

Recent genetic and archaeological evidence suggests the existence of at least four independent centres of domestication in the early Holocene epoch, two in the Old World (Near East and China) and two in the New World (southwestern Mexico and northwestern South America)¹. However, the closest wild ancestors of several globally important domesticated cultigens occur in southwestern Amazonia. These include *Manihot esculenta* subsp. *flabellifolia*, the wild ancestor of manioc (*Manihot esculenta*)⁷; *Cucurbita maxima* subsp. *andreana*, the wild ancestor of the squash (*Cucurbita maxima* subsp. *maxima*)⁸; peach palm (*Bactris gasipaes*)⁹; *Canavalia piperi*, the wild ancestor of jack bean (*Canavalia plagioperma*)⁴; and *Capsicum baccatum* var. *baccatum*, the wild ancestor of chili peppers (*Capsicum baccatum* var. *pendulum*)¹⁰. This suggests that southwestern Amazonia could be a fifth early Holocene centre of domestication. However, with the exception of *Calathea* sp. phytoliths that possibly represent *lerén* (*Calathea alluioia*) (which have recently been documented in the upper Madeira basin¹¹), archaeological evidence has not been found for early plant cultivation in southwestern Amazonia. Our research fills this gap with data from 61 archaeological sites—which we refer to as ‘forest islands’^{12–14}, because they now occur as patches of forest surrounded by savannah—dated to the early and mid-Holocene epoch.

Mapping of forest islands

We used remote-sensing data to map 6,643 forest islands in the Llanos de Moxos. The average size of forest islands is 0.5 ha (minimum of 0.05 ha and maximum of 16 ha; s.d. 0.65 ha). We surveyed 82 of these forest islands, which represents about 1.2% of all sites. We took column sediment samples

from all of the surveyed sites, and carried out archaeological excavations in four. We classified 64 out of 83 (the 82 sites we sampled, plus Monte Castelo in Brazil¹⁴) sites as anthropic on the basis of the presence of deep dark sediments rich in organic matter, charcoal and burned earth that were frequently associated with shell and bone fragments (Fig. 1). The forest islands that we surveyed are between about 0.5 m and 3 m high. The weighted proportion of anthropic versus natural sites suggests the existence of at least 4,700 anthropic forest islands in the Llanos de Moxos (Extended Data Fig. 1). This is probably far fewer than the original number built in the early and mid-Holocene epoch, as during the transition to the late Holocene epoch most of the rivers in the southwestern part of the Llanos de Moxos became very active; many of the pre-existing soils and potential archaeological sites were covered by alluvial deposits—sometimes up to 5-m thick¹⁵. This explains the modern distribution of forest islands and why 48% of the 6,643 forest islands we mapped are concentrated in a relatively small area in the northwestern Llanos de Moxos (Extended Data Fig. 1), where the landscape did not change notably during this period. Most of the anthropic forest islands are located in interfluvial settings covered by seasonally flooded savannahs; they account for an estimated 24 km² of forested area and, in aggregate, their circumferences comprise around 1,000 km of forest–savannah ecotone.

Sixty-six accelerated mass spectrometry ¹⁴C dates from 31 archaeological sites (Supplementary Table 1) bracket the human occupation of forest islands throughout the Holocene epoch to between about 10,850 and 2,300 cal. yr BP. The dated sites—except for three sites in the northeastern Llanos de Moxos, two of which are dated to 2,350 cal. yr BP and one to 4,100 cal. yr BP—were established between the early and mid-Holocene epoch.

¹Institute of Geography, University of Bern, Bern, Switzerland. ²Department of Archaeology, College of Humanities, University of Exeter, Exeter, UK. ³Laboratório de Arqueologia dos Trópicos, Museu de Arqueologia e Etnologia, Universidade de São Paulo, São Paulo, Brazil. ⁴CaSEs – Culture and Socio-Ecological Dynamics Research Group, Pompeu Fabra University, Barcelona, Spain. ⁵Department of Anthropology, The Pennsylvania State University, University Park, PA, USA. ⁶Instituto de Investigaciones Antropológicas y Arqueológicas, Universidad Mayor de San Andrés, La Paz, Bolivia. ✉e-mail: umberto.lombardo@giub.unibe.ch

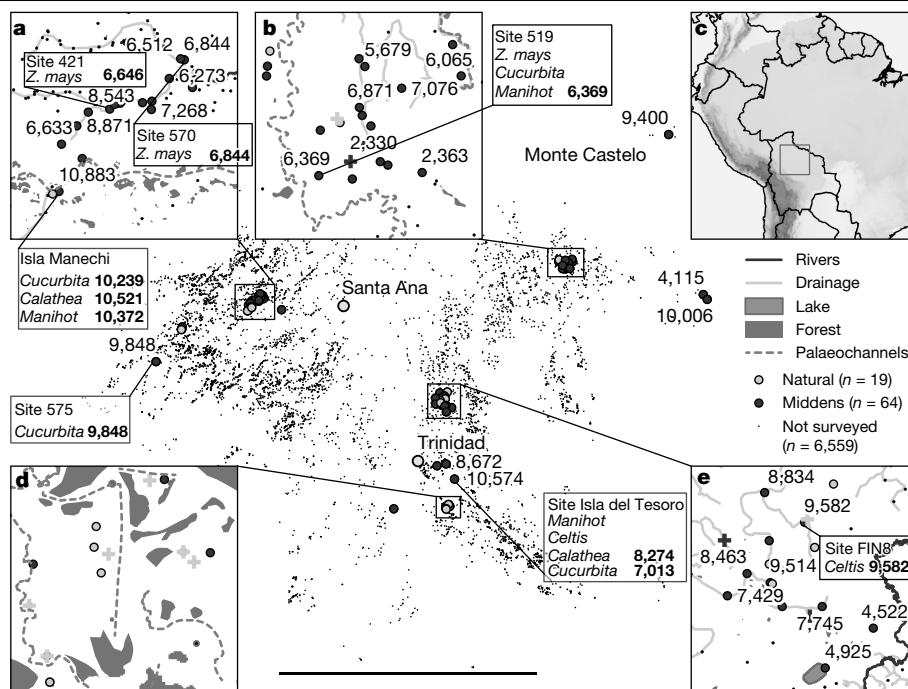


Fig. 1 | Forest islands mapped in the Llanos de Moxos. The numbers associated with middens are dates expressed in median cal. yr BP from the deepest anthropic datable layer at each site (Extended Data Table 1). **a, b, d, e.** Areas that were surveyed to estimate the total number of anthropic forest islands in the

Llanos de Moxos (Extended Data Fig. 1). **c.** Large-scale map, identifying the study area (square) and Greater Amazonia (grey shaded area). The Andes is shown in dark grey. Circles, round forest islands; crosses, irregular forest islands (see 'Mapping of forest islands' in the Methods for more detail). Scale bar, 200 km.

Evidence of early plant cultivation

We analysed phytoliths from the radiocarbon-dated profiles of 30 forest islands (Figs. 2, 3). The earliest evidence of *Manihot* is a heart-shaped phytolith¹⁶ found at the Isla Manechi that is dated to about 10,350 cal. yr BP; another phytolith from the Isla del Tesoro was dated to about 8,250 cal. yr BP (Fig. 1). Scalloped, spherical phytoliths that derive from the rind of *Cucurbita* sp. were identified in layers dated to about 10,250 cal. yr BP (Isla Manechi) and about 9,850 cal. yr BP (site 575). We identified wavy-top rondel phytoliths, which are produced in the cob of maize¹⁷, dated to about 6,850 cal. yr BP (site 570) and about 6,700 cal. yr BP (site 421). We detected the early presence of phytoliths from the rhizome of *Calathea* sp. (a member of the Marantaceae) at around 8,275 cal. yr BP (Isla del Tesoro), before about 7,800 cal. yr BP (site 433) and at approximately 7,400 cal. yr BP (site FIN14). Other phytoliths—from *Phenakospermum guyanense*, *Heliconia* sp. as well as members of the Marantaceae and Cyperaceae—have been found in almost all of the samples, starting from around 10,400 cal. yr BP (Fig. 2, Extended Data Fig. 2). We identified phytoliths derived from the seeds of *Oryza* sp. that dated to about 6,250 cal. yr BP (San Pablo), as well as phytoliths from the epidermis of seeds of *Celtis* sp. from contexts dated to around 9,600 cal. yr BP (site FIN8). Hat-shaped phytoliths, which are diagnostic of the Arecoideae subfamily¹⁷ of the Arecaceae (palms), are present at about 9,975 cal. yr BP (site 493). Peach palm (*B. gasipaes*), which is a member of this subfamily, is the only palm to have been domesticated in South America, and its domestication probably took place in southwestern Amazonia⁶. *Bactris* palms produce hat-shaped phytoliths, but do not produce phytoliths that are diagnostic at the species level¹⁸. *Bactris* spp., and other arecoid genera that grow today in the Llanos de Moxos (*Astrocaryum*, *Desmoncus*, *Geonoma* and *Socratea*), are used for food, building materials and medicine throughout present-day Amazonia¹⁹. The size of the squash phytoliths that we recovered is well within the range of those of the domesticated species (Extended Data Table 2), however, these phytoliths do not show an increase in size over time, as would be expected for a species under domestication pressure¹⁷ (Extended Data Table 2). The presence of domesticated *Cucurbita* sp. beginning at

around 10,250 cal. yr BP (Isla Manechi) is—to our knowledge—the oldest evidence for *Cucurbita* sp. in association with human activity in Amazonia, and coincides with the domestication of several species of *Cucurbita* across Central²⁰ and South America^{21,22} at the very beginning of the Holocene epoch. Further studies that analyse larger sample sizes are required to determine whether the domesticated squash cultivated in the early Holocene epoch was adopted in the Llanos de Moxos from other regions or was domesticated in situ. Maize cob phytoliths were documented at site 570 at about 6,850 cal. yr BP, which represents—to our knowledge—the oldest evidence, by a few centuries, of maize cultivation in the Amazon basin. As has previously been hypothesized²³, the early maize found in this (and other) areas probably represented a partially domesticated variety that later diverged into two South American groups of fully domesticated maize varieties. This early evidence of maize phytoliths is consistent with a temporal gradient of maize dispersal that began in western Amazonia and reached the eastern Amazon by around 4,300 cal. yr BP. The early use of *Manihot* (as documented at Isla Manechi) in Llanos de Moxos began more than 10,000 years ago, which coincides with the estimated time for the molecular divergence of the domesticated species from its wild ancestor and with the current biogeography of the closest wild ancestor of manioc^{7,24}. Manioc possibly spread later to northern Peru, Colombia and Panama (where the earliest known evidence dates to 8,500 cal. yr BP, 7,000 cal. yr BP and 7,600 cal. yr BP, respectively)⁵, suggesting that the bidirectional exchange of cultivars between Amazonia and the Andes began in the early Holocene epoch. Our study shows that, as in other regions of Amazonia and Central America, in the Llanos de Moxos the development or arrival of full-blown agricultural societies was a very late phenomenon⁹; there is no evidence of land prepared for agriculture in the Llanos de Moxos until raised fields and drainage canals were built around 1,500–1,000 years ago²⁵.

The importance of starch-based foods

Palaeoecological studies indicate that the Llanos de Moxos was covered by cerrado-like savannah during the early and mid-Holocene epoch¹⁵,

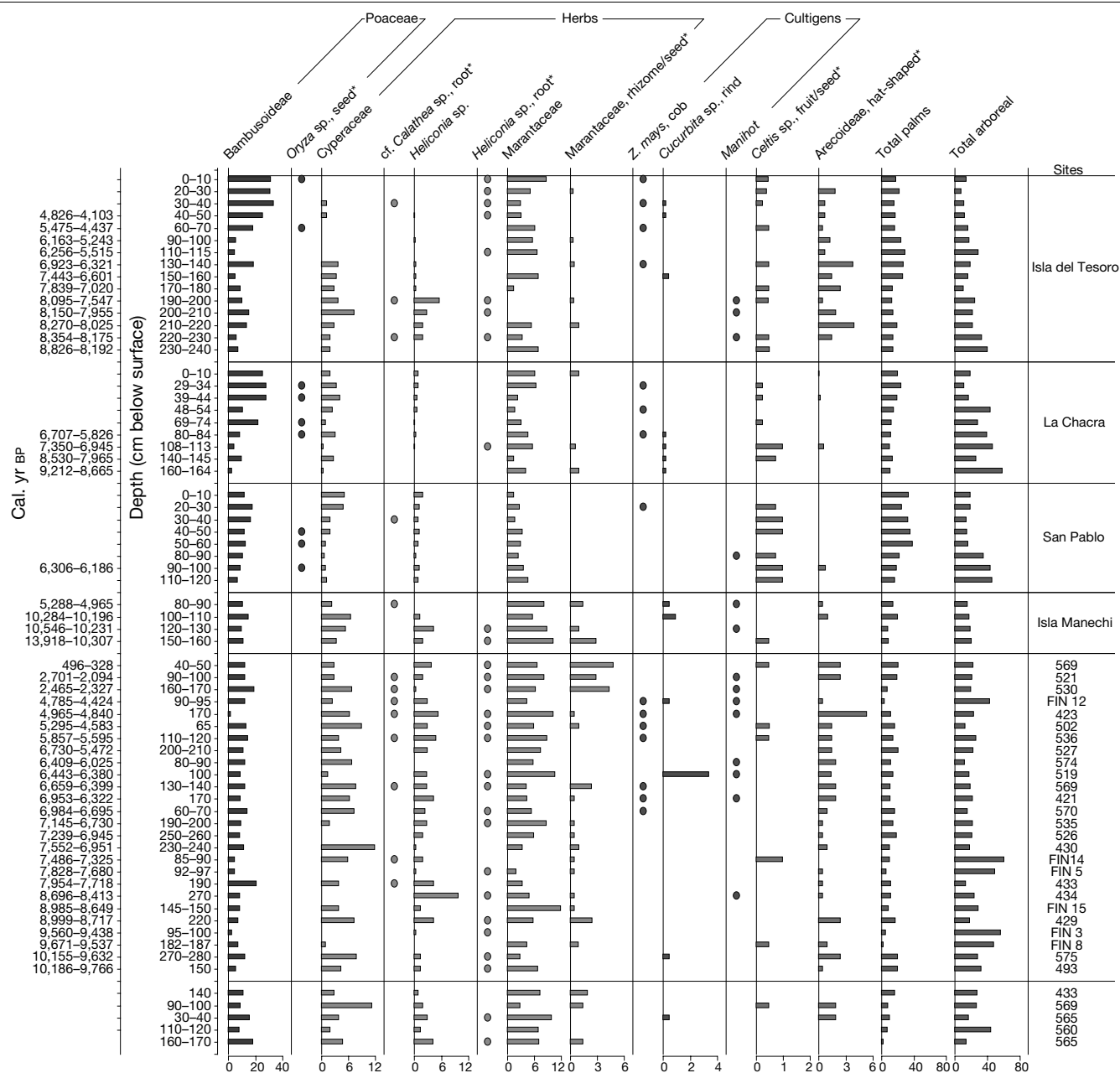


Fig. 2 | Diagram showing the percentage of the most relevant phytolith groups from anthropic forest islands. Dots indicate presence $\leq 1\%$ (for more details, see Extended Data Fig. 2, Extended Data Table 1). Asterisks denote

non-domesticated edible plants that are used as food resources today³⁵. Phytolith percentages are based on a minimum of 200 diagnostic phytoliths counted per slide.

indicating that plant cultivation in the Neotropics started in shrub savannahs as well as seasonal tropical forest environments⁴. It does not come as a surprise that phytoliths derived from plants that produce underground storage organs—including *Manihot*, *Calathea* and other *Marantaceae*, *Heliconia*, *Cyperaceae* and *Phenakospermum*—constitute an important part of the total phytolith assemblages from the forest islands of the Llanos de Moxos. These plants are abundant in savannahs and produce carbohydrate-rich foods that, with the exception of some varieties of manioc²⁶, are easy to process and cook. Today, they are consumed by indigenous groups¹⁹ and they probably provided a considerable part of the calories consumed by the first inhabitants of the Llanos de Moxos. The large herbivores and fish available in the savannahs^{12,13} would have complemented a mixed economy. The fertile forest islands were probably the home gardens in which these crops were cultivated. Our data are consistent with the hypothesis that plants that produce underground storage organs were a fundamental part of the diet of human populations as they colonized new territories^{27,28}.

Implications for biodiversity

Our results show that inland savannahs were a key region for the early occupation of the Neotropics, and that these savannahs began to be transformed by the arrival of very early human settlers. Anthropic forest islands are entirely artificial, and do not take advantage of pre-existing landscape features. Their formation is not only an incidental effect of food-waste dumping, but can also be seen as an active process of niche construction²⁹. These accumulative middens constituted fertility hotspots amid poor savannah soils, because (i) they were loci for the accumulation of nutrients that came from gathering activities in the surrounding savannah and (ii) they remained above the water level during the wet season¹². It is only after 4,000 years BP, when the old and infertile soils of the south of the Llanos de Moxos were covered with fertile alluvium deposited by the Río Grande, that agriculture in the savannahs was facilitated³⁰. Overall, the early-to-mid-Holocene

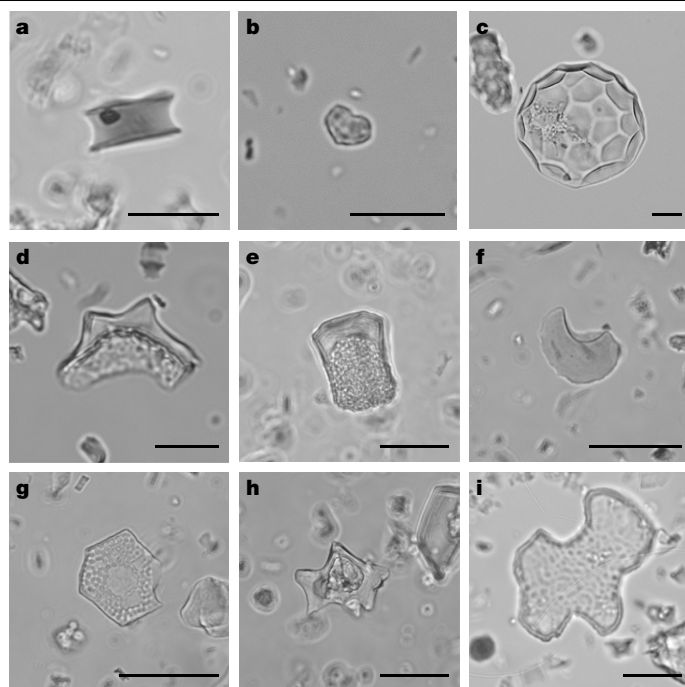


Fig. 3 | Photomicrographs of phytolith morphotypes recovered from Isla del Tesoro, La Chacra and Isla Manechi. a, Wavy-top rondel from the cob of maize (*Z. mays*) (sample code IT190-200). **b**, Heart-shaped phytolith from the secretory cells of manioc (*Manihot*) (sample code BANR17-UE1-57). **c**, Scalloped sphere from the rind of squash (*Cucurbita* sp.) (sample code BANR17-UE1-31). **d**, Double-peaked glume from the seed of rice (*Oryza* sp.) (sample code SM3-116s). **e**, Flat domed cylinder from the rhizome of *Calathea* sp. (sample code IT30-40). **f**, Short trough body from the rhizome of *Heliconia* sp. (sample code IT130-140). **g**, Stippled polygonal body from the seed of a member of the Cyperaceae (sample code IT150-170). **h**, Phytolith with nodular projections and a pointed apex, from the seed of Marantaceae (sample code IT90-100). **i**, Stippled plate from the fruit of a hackberry (*Celtis* sp.) (sample code SM3-69-74). Scale bars, 20 μ m.

construction of forest islands—which became key structures in the landscape^{31,32}—increased forest patchiness (Extended Data Figs. 3, 4a) and probably contributed to maintaining landscape-scale species richness in this threatened biome, which is a wetland designated under the Ramsar Convention (<https://whc.unesco.org/en/ramsar/>). Nowadays, these anthropic forest islands are preferential feeding and roosting sites for many species of birds, including the endemic and critically endangered blue-throated macaw (*Ara glaucogularis*)³³. Taken together, our data show that the earliest inhabitants of the Llanos de Moxos relied not only on foraging but had also engaged in plant cultivation since the early Holocene epoch, thus opening up the possibility that they already had a mixed economy when they arrived in the region. The thousands of keystone structures represented by forest islands show that the human footprint on Amazonia is not restricted to large-scale transformations by farming groups in late Holocene epoch^{9,34}, but is instead rooted in the earliest human dispersal into this region—and has lasting implications for habitat heterogeneity and biodiversity conservation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2162-7>.

- Larson, G. et al. Current perspectives and the future of domestication studies. *Proc. Natl Acad. Sci. USA* **111**, 6139–6146 (2014).
- Zohary, D. & Hopf, M. *Domestication of Plants in the Old World: the Origin and Spread of Cultivated Plants in West Asia, Europe and the Nile Valley* (Oxford Univ. Press, 2000).
- Zeder, M. A., Bradley, D. G., Smith, B. D. & Emshwiller, E. *Documenting Domestication: New Genetic and Archaeological Paradigms* (Univ. California Press, 2006).
- Piperno, D. R. & Pearsall, D. M. *The Origins of Agriculture in the Lowland Neotropics* (Academic, 1998).
- Piperno, D. R. The origins of plant cultivation and domestication in the New World tropics: patterns, process, and new developments. *Curr. Anthropol.* **52**, S453–S470 (2011).
- Clement, C. R., de Cristo-Araújo, M., d'Eeckenbrugge, G. C., Alves Pereira, A. & Picanço-Rodrigues, D. Origin and domestication of native Amazonian crops. *Diversity (Basel)* **2**, 72–106 (2010).
- Olsen, K. & Schaal, B. Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. *Am. J. Bot.* **88**, 131–142 (2001).
- Sanjurjo, O. I., Piperno, D. R., Andres, T. C. & Wessel-Beaver, L. Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: implications for crop plant evolution and areas of origin. *Proc. Natl Acad. Sci. USA* **99**, 535–540 (2002).
- Clement, C. R. et al. The domestication of Amazonia before European conquest. *Proc. R. Soc. Lond. B* **282**, 20150813 (2015).
- Scaladaferro, M. A., Barboza, G. E. & Acosta, M. C. Evolutionary history of the chili pepper *Capsicum baccatum* L. (Solanaceae): domestication in South America and natural diversification in the seasonally dry tropical forests. *Biol. J. Linn. Soc.* **124**, 466–478 (2018).
- Watling, J. et al. Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. *PLoS ONE* **13**, e0199868 (2018).
- Lombardo, U. et al. Early and middle Holocene hunter-gatherer occupations in western Amazonia: the hidden shell middens. *PLoS ONE* **8**, e72746 (2013).
- Capriles, J. M. et al. Persistent Early to Middle Holocene tropical foraging in southwestern Amazonia. *Sci. Adv.* **5**, eaav5449 (2019).
- Hilbert, L. et al. Evidence for mid-Holocene rice domestication in the Americas. *Nat. Ecol. Evol.* **1**, 1693–1698 (2017).
- Lombardo, U. et al. Holocene land cover change in south-western Amazonia inferred from paleoflood archives. *Global Planet. Change* **174**, 105–114 (2019).
- Chandler-Ezell, K., Pearsall, D. M. & Zeidler, J. A. Root and tuber phytoliths and starch grains document manioc (*Manihot esculenta*) arrowroot (*Maranta arundinacea*) and llerén (*Calathea* sp.) at the Real Alto site, Ecuador. *Econ. Bot.* **60**, 103–120 (2006).
- Piperno, D. R. *Phytoliths* (AltaMira Press, 2006).
- Morote-Rios, G., Bernal, R. & Raz, L. Phytoliths as a tool for archaeobotanical, palaeobotanical and palaeoecological studies in Amazonian palms. *Bot. J. Linn. Soc.* **182**, 348–360 (2016).
- Hanelt, P., Buttner, R. & Mansfeld, R. *Mansfeld's Encyclopedia of Agricultural and Horticultural Crops (except Ornamentals)* (Springer, 2001).
- Smith, B. D. The initial domestication of *Cucurbita pepo* in the Americas 10,000 years ago. *Science* **276**, 932–934 (1997).
- Piperno, D. R. & Stothert, K. E. Phytolith evidence for early Holocene *Cucurbita* domestication in southwest Ecuador. *Science* **299**, 1054–1057 (2003).
- Dillehay, T. D. & Piperno, D. R. in *The Cambridge World Prehistory* (eds Renfrew, C. & Bahn, P.) 970–985 (Cambridge Univ. Press, 2014).
- Kistler, L. et al. Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* **362**, 1309–1313 (2018).
- Rival, L. & McKey, D. Domestication and diversity in manioc (*Manihot esculenta* Crantz ssp. *esculenta*, Euphorbiaceae). *Curr. Anthropol.* **49**, 1119–1128 (2008).
- Rodrigues, L., Lombardo, U. & Veit, H. Design of pre-Columbian raised fields in the Llanos de Moxos, Bolivian Amazon: differential adaptations to the local environment? *J. Archaeol. Sci. Rep.* **17**, 366–378 (2018).
- McKey, D., Cavagnaro, T. R., Cliff, J. & Gleadow, R. J. C. Chemical ecology in coupled human and natural systems: people, manioc, multitrophic interactions and global change. *Chemoecology* **20**, 109–133 (2010).
- Jones, M. in *The Evolution of Hominin Diets* (eds Hublin, J.-J. & Richards, M. P.) 171–180 (Springer, 2009).
- Aceituno, F. J. & Loaiza, N. The origins and early development of plant food production and farming in Colombian tropical forests. *J. Anthropol. Archaeol.* **49**, 161–172 (2018).
- Smith, B. D. General patterns of niche construction and the management of 'wild' plant and animal resources by small-scale pre-industrial societies. *Phil. Trans. R. Soc. Lond. B* **366**, 836–848 (2011).
- Lombardo, U., May, J.-H. & Veit, H. Mid- to late-Holocene fluvial activity behind pre-Columbian social complexity in the southwestern Amazon basin. *Holocene* **22**, 1035–1045 (2012).
- Manning, A. D., Fischer, J. & Lindenmayer, D. B. Scattered trees are keystone structures – implications for conservation. *Biol. Conserv.* **132**, 311–321 (2006).
- Tews, J. et al. Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures. *J. Biogeogr.* **31**, 79–92 (2004).
- Berkunsky, I. et al. Assessing the use of forest islands by parrot species in a Neotropical savanna. *Avian Conserv. Ecol.* **10**, 11 (2015).
- Prümers, H. & Jaimes Betancourt, C. 100 años de investigación arqueológica en los Llanos de Moxos. *Arqueoantropológicas* **4**, 11–53 (2014).
- Junqueira, A. B., Shepard, G. H. & Clement, C. R. J. E. B. Secondary forests on anthropogenic soils of the middle Madeira river: valuation, local knowledge, and landscape domestication in Brazilian Amazonia. *Econ. Bot.* **65**, 85–99 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Standard sample size for phytolith studies of 200 diagnostic phytoliths was used. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Mapping of forest islands

Forest islands were mapped by visual scanning of the high-resolution satellite imagery of Esri ArcGIS base maps (Extended Data Fig. 5). When the identification of forest islands was not straightforward (owing to cloud cover or poor resolution), TanDemX and SRTM (Shuttle Radar Topography Mission) digital elevation models were used as complementary resources. Patches of forest were classified as forest islands when they had a round shape and were completely or partially surrounded by savannah ($n = 4,339$); or they had an irregular shape but were relatively small (<400 m in diameter) and completely surrounded by savannah ($n = 2,304$). For each forest island, the following attributes were recorded: diameter; shape (perfectly round, almost round, elongated or irregular); location (along a palaeochannel, along a modern river, seasonally flooded savannah, border between seasonally flooded savannah and upland, along a drainage stream, or upland surrounded by bushes); presence of other earthworks within about 500 m; and whether or not forest islands were established over fluvial deposits or uplands. The latter attribute is partly redundant with location, but sometimes fluvial deposits are not connected to palaeochannels (as in the case of old crevasse splays or old meander belts in which palaeochannels have been infilled) or the forest islands are located along a palaeochannel with completely eroded levees and the forest islands have clearly been built after the erosion of the levees.

Selection of survey areas

Four survey areas were selected in different regions of the Llanos de Moxos to ground-truth the forest islands identified from remote-sensing imagery, and evaluate their natural or anthropic origin. The four areas (Fig. 1) were selected on the basis of differences in soil, landcover, hydrology and accessibility by car. These four areas cover all of the different eco-regions identified in the Llanos de Moxos^{36,37}. These areas belonged to organizations (as with area a (shown in Fig. 1a), which is in the Barba Azul nature reserve) or ranchers (as with areas b, d and e (shown in Fig. 1b, d, e, respectively)) who granted us permission to conduct surveys. In total, we surveyed 21 forest islands in area a, 22 in area b, 13 in area d, and 17 in area e. Nine other forest islands, found outside of these four areas, were surveyed.

Criteria for the identification of anthropic forest islands

Several anthropic forest islands in the Llanos de Moxos have previously been excavated^{12–14}. These excavations have revealed thick strata of sediments rich in organic matter, charcoal, burnt earth and fragmented animal bones and shells; they also have revealed human burials. The clear difference between the sediments found in the anthropic forest islands and the soil types found in the Llanos de Moxos^{38–41} makes the field identification of forest islands relatively straightforward. In the present work, the forest islands surveyed have been classified as anthropic when thick layers of organic-rich sediments contained at least two archaeological materials (such as charcoal, burnt earth, animal bones or shells).

Sampling of forest islands

Sampling of undisturbed material was performed at regular intervals in the four sites at which archaeological excavations were conducted: Isla del Tesoro (site code SMI), La Chacra (site code SM3), San Pablo (site code SM4) and Isla Manechi (Extended Data Fig. 6). The rest of the sites were sampled using an auger soil sampler. The stratigraphy of the recovered cores was described in the field, and sampling was carried out only where stratigraphic changes were detected in the field

(Extended Data Fig. 7). The deepest sample with evidence of charcoal was always sampled. After extraction, cores were inspected to avoid contamination and check that the soil section that we sampled showed no evidence of soil mixing (that is, root penetration and invertebrate burrowing were absent). The excess of material was cut off with a knife and only the inner, uncontaminated part of the extracted samples was been stored in plastic bags. Samples were air-dried in Bolivia before being shipped. Charcoal fragments for ¹⁴C dating were collected in situ, enveloped in aluminium foil and stored in plastic bags.

Phytolith processing and identification

Phytoliths were extracted from sediments following previously published methods⁴². Phytoliths were identified and counted using a Zeiss Axioscope 40 light microscope at 500× magnification. Phytolith identifications were made using published material for the Neotropics^{17,43–46} and by direct comparison with the phytolith reference collection of the Archaeobotany and Palaeoecology Laboratory (Department of Archaeology, University of Exeter). A minimum of 200 diagnostic phytoliths were counted per slide. A full scan of the slides was performed to detect the presence of squash, manioc and maize. Phytolith assemblages in southwestern Amazonia have been studied in modern soils⁴⁶ and 29 palaeosols from the early and late Holocene epoch¹⁵ from different natural environments and land covers. Phytoliths of *Manihot* or *Curcubita* have not been found in any of these natural contexts, which strongly suggests that the phytoliths of these two genera found in forest islands are the direct result of human activity and not of the chance occurrence of wild relatives on the forest islands.

Radiocarbon dates

The deepest recoverable sample of charcoal from 32 sites was dated to establish the minimum date for the foundation of the site. For the 4 sites that were excavated, 35 samples from different depths were dated to establish periods of occupation and abandonment. The complete dataset and code used to calibrate all of radiocarbon dates are available in Extended Data Table 1 and the Supplementary Information, respectively. Radiocarbon dates from the studied sites were calibrated using SHCAL13⁴⁷. For Isla del Tesoro, Isla Manechi and La Chacra (for which stratigraphically ordered ages were available), we ran a series of Bayesian age–depth models using the P_Sequence command in OxCal 4.3⁴⁸ with default settings. Each model was stratigraphically constrained by the youngest age in the profile and the deepest section reached in each site (Extended Data Fig. 8). The ages of the undated samples were estimated using the command Date within the model.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All relevant data, including Source Data for Figs. 1, 2, are provided with the paper.

Code availability

Code used for the calibration of the ¹⁴C dates in OxCal is available in the Supplementary Information.

36. Lombardo, U., Canal-Beeby, E. & Veit, H. Eco-archaeological regions in the Bolivian Amazon: linking pre-Columbian earthworks and environmental diversity. *Geogr. Helv.* **66**, 173–182 (2011).
37. Langstroth Plotkin, R. Biogeography of the Llanos de Moxos: natural and anthropogenic determinants. *Geogr. Helv.* **66**, 183–192 (2011).
38. Lombardo, U., Denier, S. & Veit, H. Soil properties and pre-Columbian settlement patterns in the monumental mounds region of the Llanos de Moxos, Bolivian Amazon. *Soil (Göttingen)* **1**, 65–81 (2015).

39. Rodrigues, L., Lombardo, U., Canal Beeby, E. & Veit, H. Linking soil properties and pre-Columbian agricultural strategies in the Bolivian lowlands: the case of raised fields in Exaltación. *Quat. Int.* **437**, 143–155 (2017).
40. Boixadera, J., Poch, R. M., García-González, M. T. & Vizcayno, C. Hydromorphic and clay-related processes in soils from the Llanos de Moxos (northern Bolivia). *Catena* **54**, 403–424 (2003).
41. Hanagarth, W. *Acerca de la Geoecología de las Sabanas del Beni en el Noreste de Bolivia* (Instituto de Ecología, 1993).
42. Lombardo, U., Ruiz-Pérez, J. & Madella, M. Sonication improves the efficiency, efficacy and safety of phytolith extraction. *Rev. Palaeobot. Palynol.* **235**, 1–5 (2016).
43. Piperno, D. R. Identifying crop plants with phytoliths (and starch grains) in Central and South America: a review and an update of the evidence. *Quat. Int.* **193**, 146–159 (2009).
44. Iriarte, J. Assessing the feasibility of identifying maize through the analysis of cross-shaped size and three-dimensional morphology of phytoliths in the grasslands of southeastern South America. *J. Archaeol. Sci.* **30**, 1085–1094 (2003).
45. Watling, J. et al. Differentiation of Neotropical ecosystems by modern soil phytolith assemblages and its implications for palaeoenvironmental and archaeological reconstructions II: southwestern Amazonian forests. *Rev. Palaeobot. Palynol.* **226**, 30–43 (2016).
46. Dickau, R. et al. Differentiation of Neotropical ecosystems by modern soil phytolith assemblages and its implications for palaeoenvironmental and archaeological reconstructions. *Rev. Palaeobot. Palynol.* **193**, 15–37 (2013).
47. Hogg, A. G. et al. SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP. *Radiocarbon* **55**, 1889–1903 (2013).
48. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).
49. Piperno, D. R. et al. Phytoliths in *Cucurbita* and other Neotropical Cucurbitaceae and their occurrence in early Archaeological sites from the lowland American tropics. *J. Arch. Sci.* **27**, 193–208 (2000).

Acknowledgements We acknowledge the support of the Bolivian Ministerio de Culturas y Turismo, the Gobierno Autónomo Departamental del Beni and the owners of the properties on which the study sites are located: J. P. Llapiz, F. Boheme, J. Rivero, O. Sikuajara, F. Velasco and T. Boorsma from the Barba Azul Natural Reserve. We thank L. Rodrigues, N. Zihlmann, G. P. Fernández and L. M. Ortega for participation during fieldwork, as well as E. Canal-Beeby, M. Madella, D. McKey, J. Carson, M. González and S. Tin for their support at different stages of this work. This work was supported by Swiss National Science Foundation grant numbers 200020-141277/1 and P300P2_158459/1; Marie Skłodowska-Curie Actions EU Project 703045; PAST project funded by the European Research Council (ERC), grant agreement number ERC_Cog 616179; National Geographic Society grant HJ-074ER-17; TerraSAR-X/TanDEM-X mission, grant number DEM_OTHER1040; and AHRC-FAPESP MoU research grant HERCA, reference AH/SO01662/1.

Author contributions U.L. designed the research. U.L., J.M.C., J.R.-P. and H.V. conducted the fieldwork. J.M.C. and J.R.-P. conducted the archaeological excavations, U.L. conducted the GIS mapping and analyses. U.L., L.H., J.R.-P. and J.I. carried out the phytolith analyses. U.L. and J.I. wrote the paper with the help of all authors.

Competing interests The authors declare no competing interests.

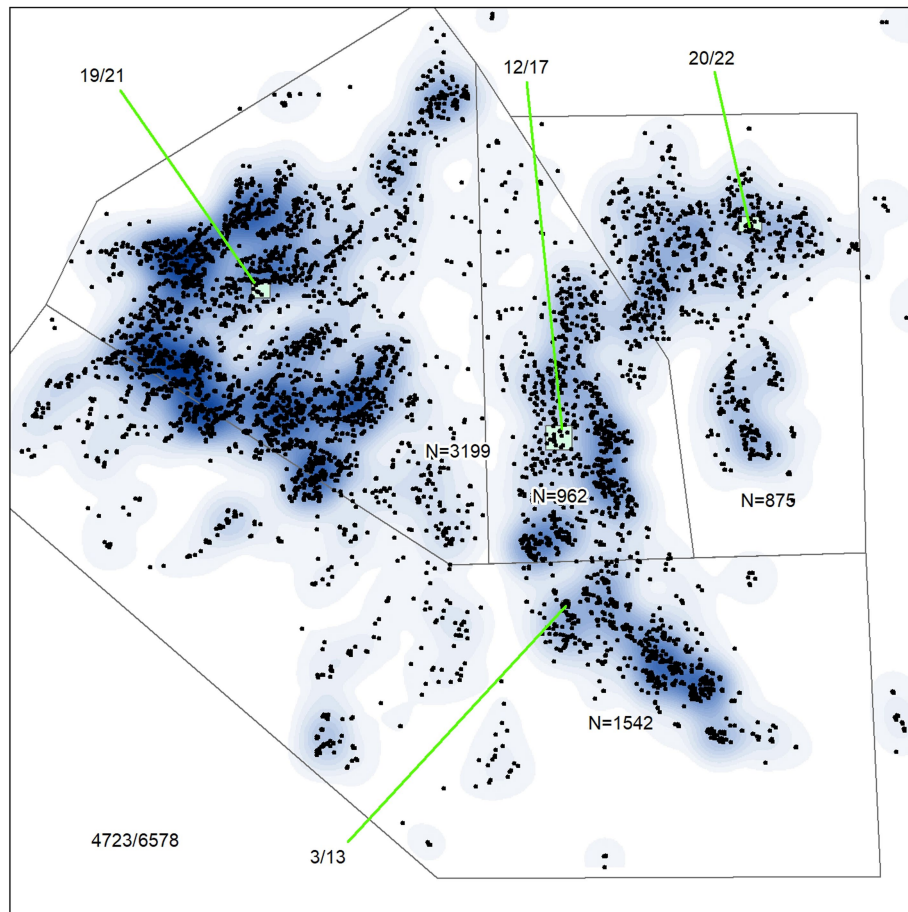
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2162-7>.

Correspondence and requests for materials should be addressed to U.L.

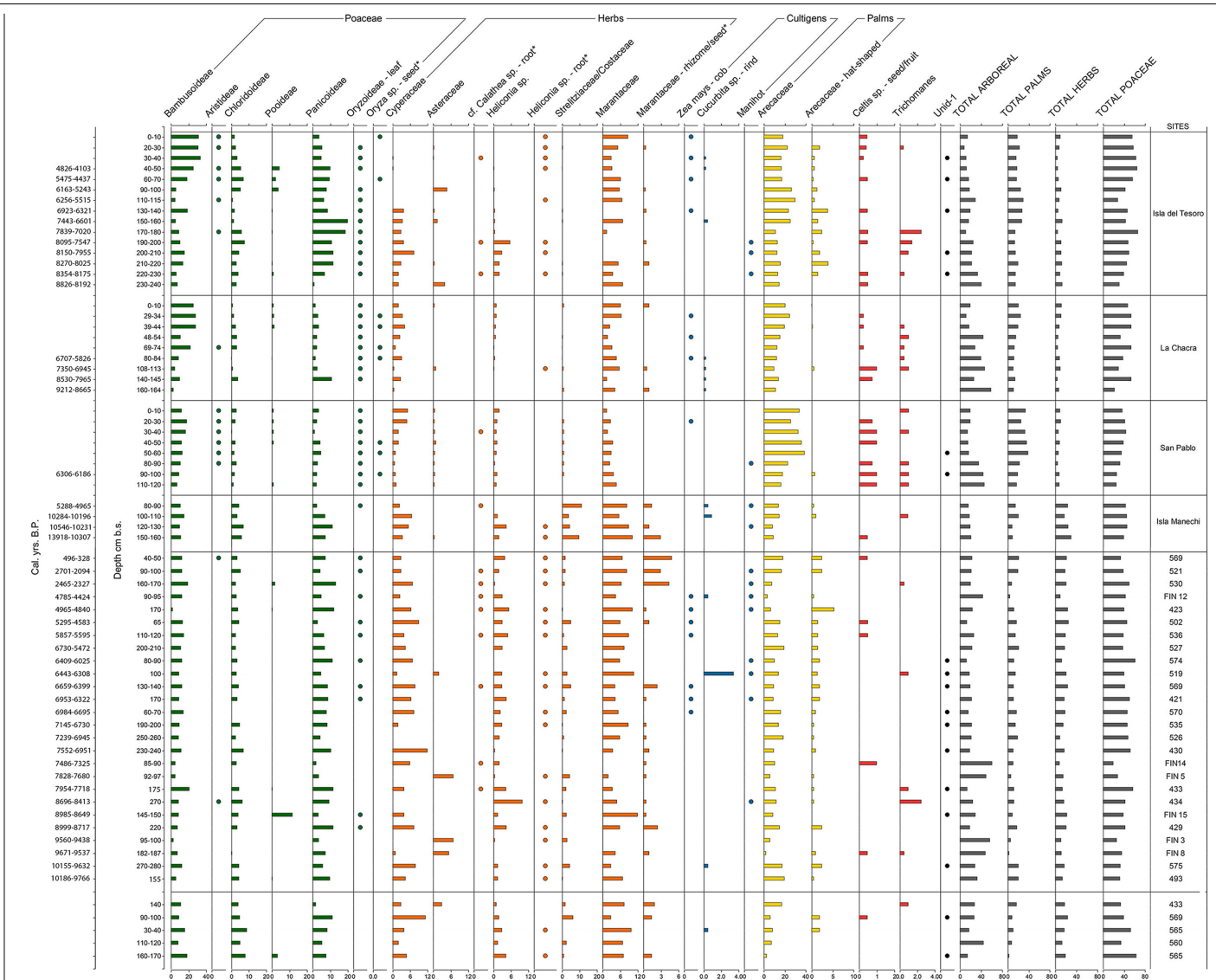
Peer review information *Nature* thanks Maarten Blaauw, Francis Mayle and Deborah M. Pearsall for their contribution to the peer review of this work.

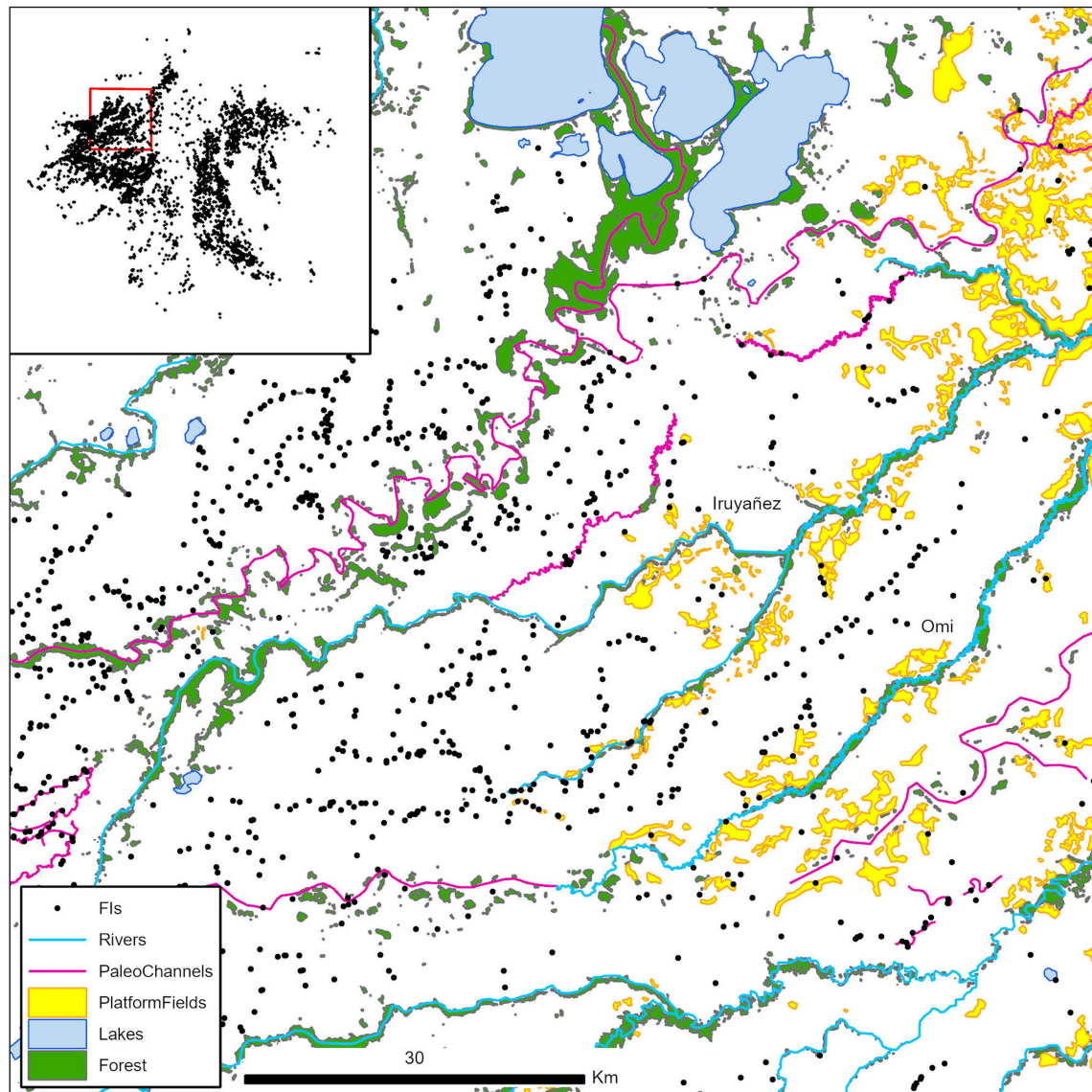
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Estimated number of anthropic forest islands. The number was estimated by extending the proportion of anthropic forest islands in the surveyed areas (in green) to the portion of the Llanos de Moxos with

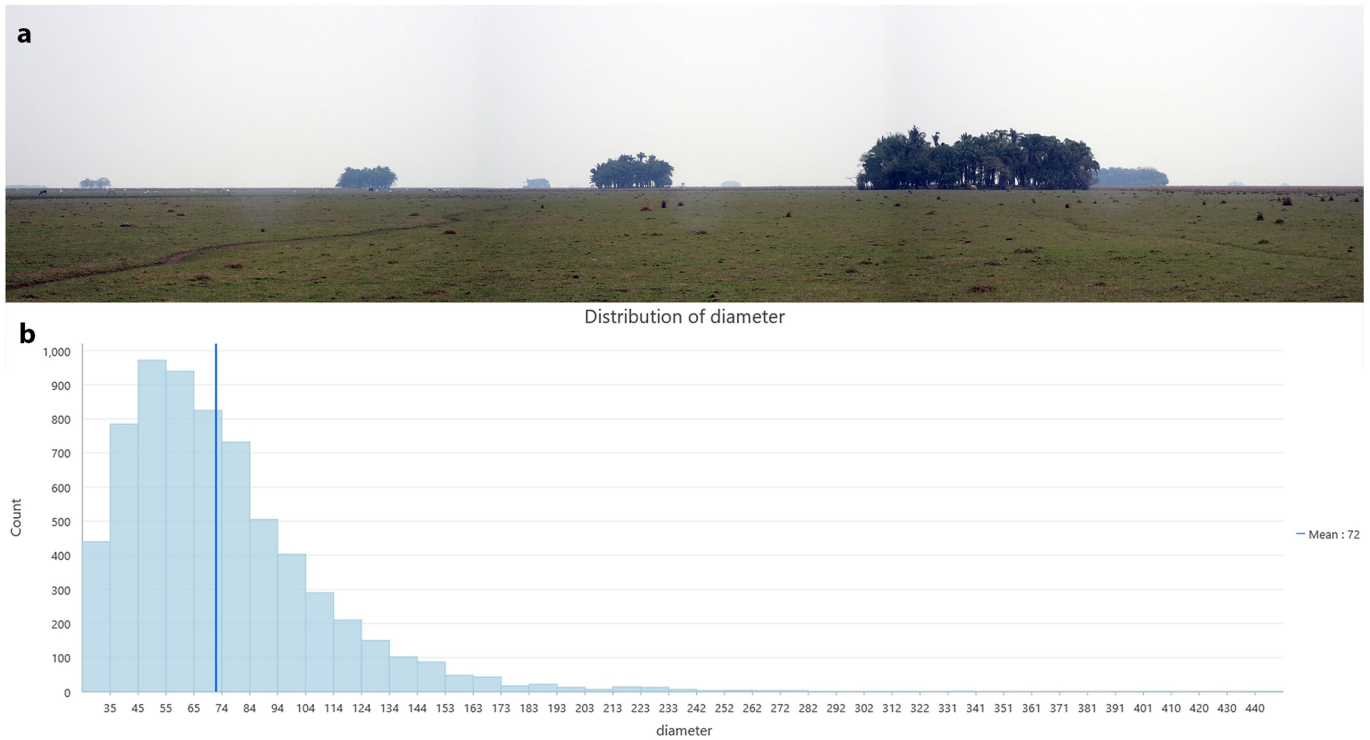
similar physical geography and land cover (polygons). Background image represents the density of forest islands, calculated using the kernel density tool in Esri ArcGIS.





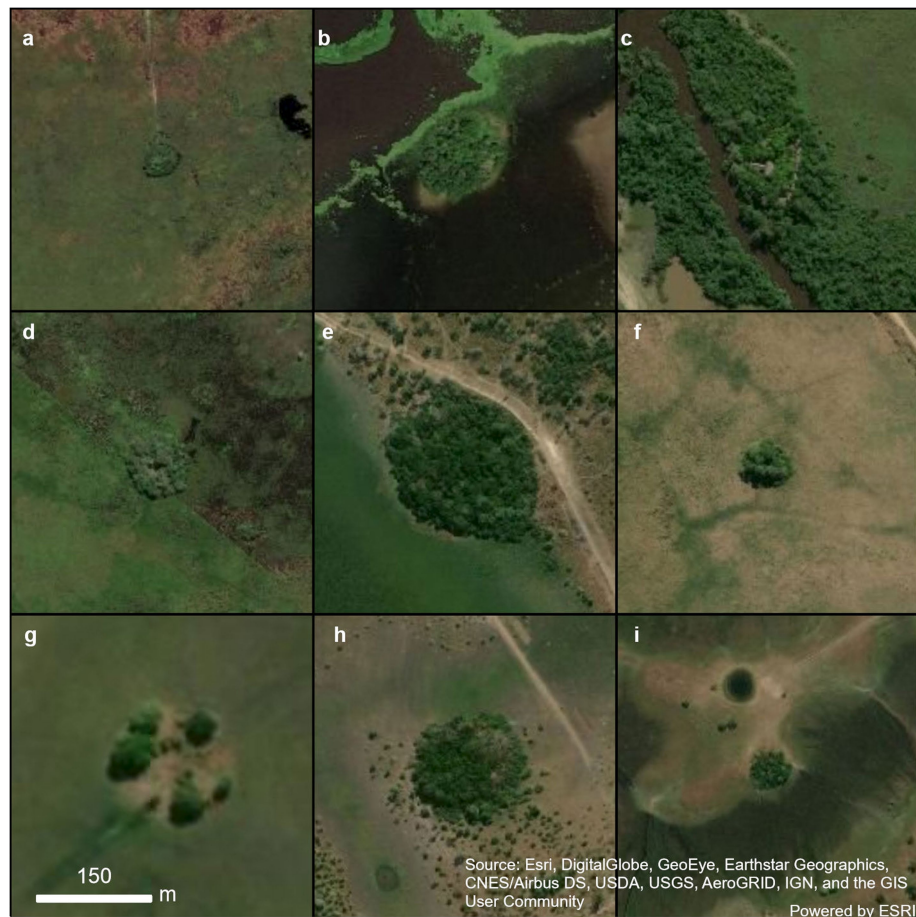
Extended Data Fig. 3 | Map of all of the forest islands and platform fields in a northwestern subset of the Llanos de Moxos. Platform fields are mostly built along palaeochannels. The Omi and Yruyañez Rivers flow inside old channels of the Beni River. Most forest islands are located in interfluvial areas. The region

contains a total of 2,428 patches of forest, 955 of which are forest islands. Once all of the patches of forest within a 2-km buffer of a river, palaeoriver and lake are removed, forest islands account for 60% of the remaining 1,191 patches.



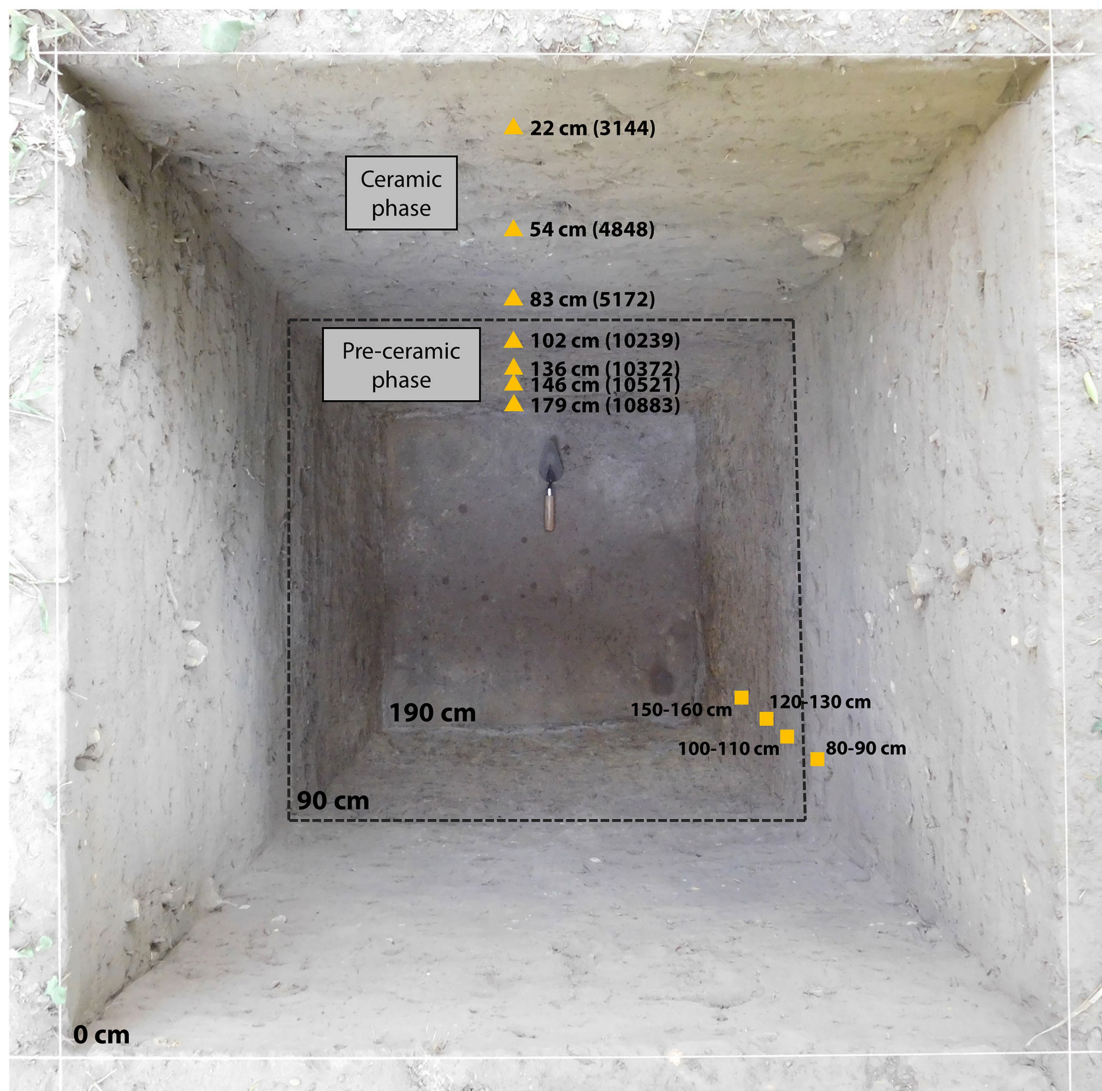
Extended Data Fig. 4 | Characteristics of forest islands. **a**, Photograph of the anthropic landscape dotted with forest islands in the Barba Azul nature reserve. **b**, Histogram showing the distribution of the diameter of forest islands. The left side of the distribution is truncated at 25 m because smaller forest islands have not been mapped. **c**, Photograph taken at site 579, a natural forest island. Samples are taken every 10 cm, from top left (highest sample) to bottom right (deepest sample). Material is silt with no organic matter.

d, Photograph taken at site 425, an anthropogenic forest island. The depth from which the sample was taken is 140 cm. **e**, Photograph taken at site 430, an anthropogenic forest island. The depth from which the sample was taken is 160 cm. Samples in **d**, **e** are representative of the whole profiles of sites 425 and 430, respectively (Extended Data Fig. 7). We obtained one core for each forest island that we visited.



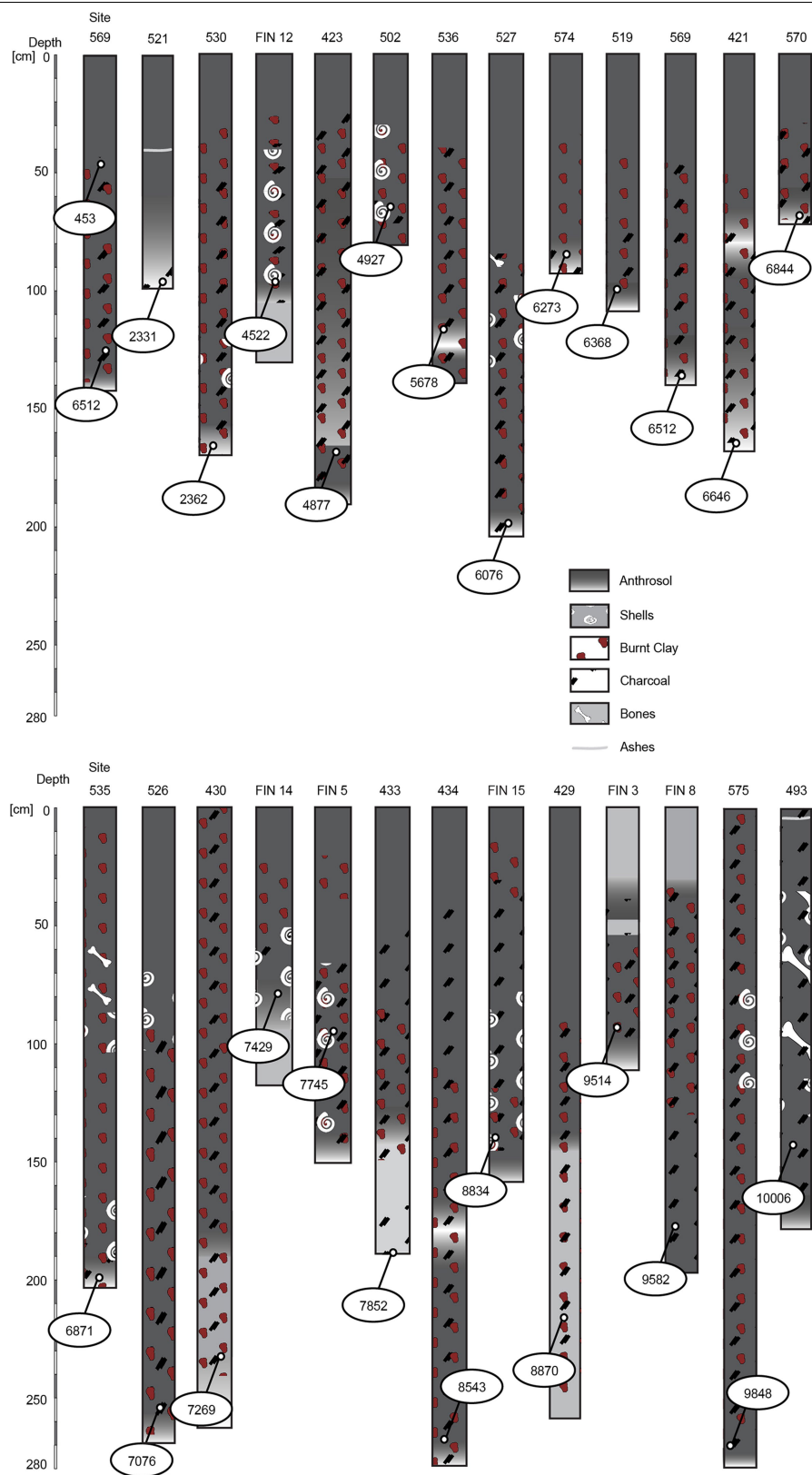
Extended Data Fig. 5 | Examples of surveyed forest islands as seen in high-resolution satellite imagery from the Esri ArcGIS base map. a–f, Forest islands classified as anthropic (**a**, Isla San Pablo (SM4); **b**, Isla Manechi; **c**, site 575; **d**, Isla La Chacra (SM3); **e**, site FIN12; and **f**, Isla del Tesoro (SM1)).

g–i, Forest islands classified as natural (**g**, site FIN2; **h**, site FIN11; and **i**, site 529). Source for the maps, ESRI, DigitalGlobe, GeoEye, Earthstar Geographics and CNES/Airbus DS.

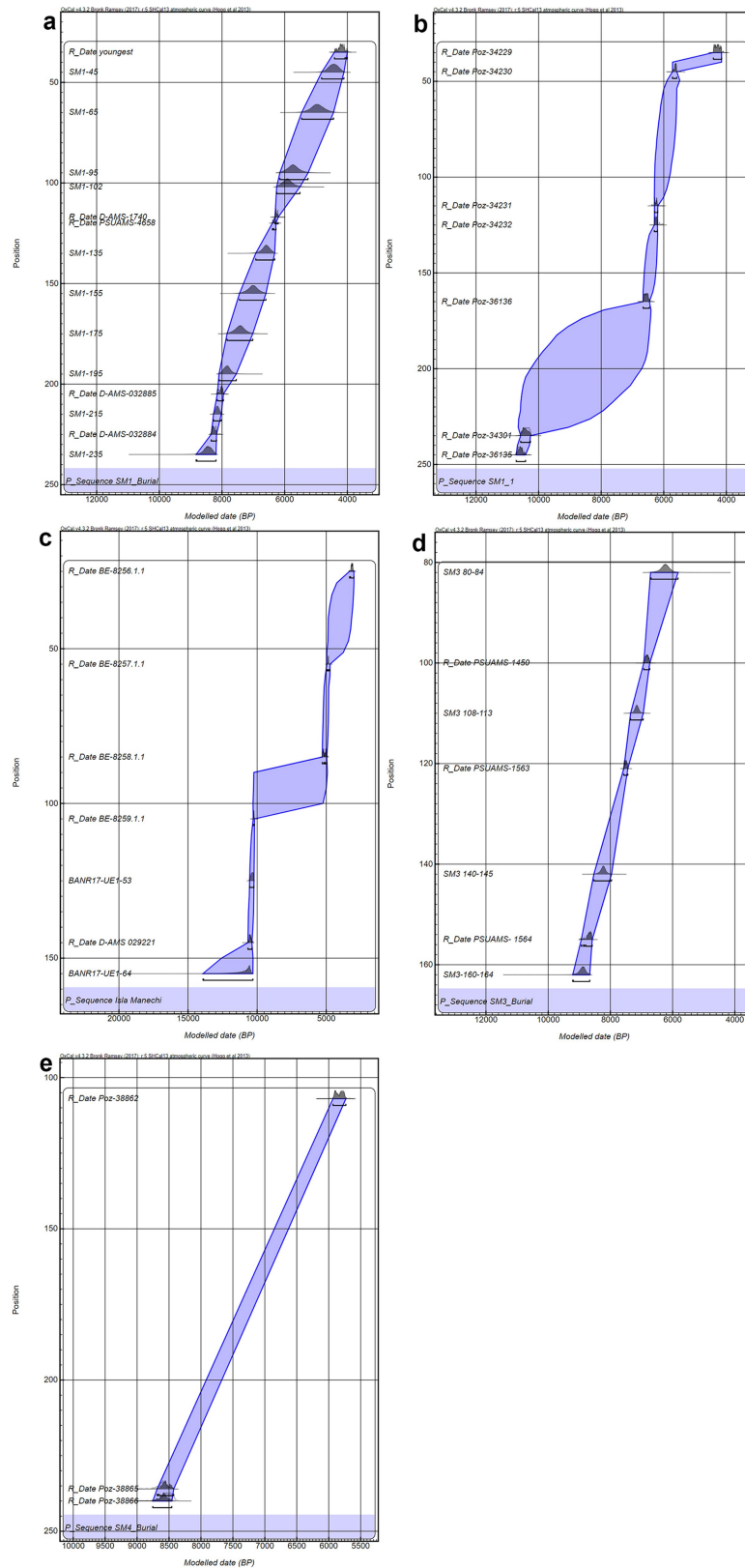


Extended Data Fig. 6 | Stratigraphic profile and sampling sites at Isla Manechi. Charcoal fragments for accelerated mass spectrometry ^{14}C dating were collected during the excavation; yellow triangles indicate their depth. Yellow squares indicate the locations of sediment samples that were analysed for phytoliths; these were sampled after the excavation from the vertical profile. The dashed line indicates the transition from the ceramic phase to the pre-ceramic phase. This transition is characterized by a sharp increase in the

compactness of sediments and amount of burnt earth. Values in parentheses are median cal. yr BP. A chronological gap of almost 4,500 years exists between the two phases (Extended Data Table 1). All of these differences make it possible to exclude any contamination of the pre-ceramic contexts with phytoliths that come from the ceramic contexts. Descriptions of the archaeological excavations at Isla del Tesoro (SM1), La Chacra (SM3) and San Pablo (SM4) have previously been published¹³.



Extended Data Fig. 7 | Stratigraphic descriptions of cored sites. Radiocarbon dates are included in ovals as median cal. yr BP.



Extended Data Fig. 8 | Age–depth models of the modelled profiles. a, b, Isla del Tesoro (SM1). **c**, Isla Manechi. **d, e**, Isla La Chacra (SM3). The age–depth models have been produced using OxCal v4.3; code is available in the Supplementary Information.

Article

Extended Data Table 1 | Radiocarbon ages of all of the dated sites cited in the text, including provenience and calibrations

Lab_Code	Site	14C BP	14C Age SD	Unmodelled (BP) 95.4%		Modelled (BP) 95.4%		Median	Material	Depth (cm)
				from	to	from	to			
D-AMS-1737	SM4, San Pablo	5476	35	6306	6124			6239	Bulk	58
Poz-46397	SM4, San Pablo	5190	80	6178	5664			5898	Charcoal	65
D-AMS-1741	SM4, San Pablo	5490	32	6306	6186			6245	Bulk	93
PSUAMS-4659	SM4, San Pablo	6665	25	7571	7441			7512	Charcoal	150
D-AMS-1739	SM4, San Pablo	6910	30	7787	7621			7696	Charcoal	150
Poz-46396	SM4, San Pablo	7700	90	8636	8218			8463	Charcoal	197
PSUAMS-1450	SM3, La Chacra	6030	30	6935	6736	6925	6734	6825	Charcoal	100
Poz-38862	SM3, La Chacra	5140	40	5930	5733	5933	5733	5825	Shell	107
PSUAMS-1563	SM3, La Chacra	6650	30	7566	7437	7568	7441	7506	Charcoal	121
PSUAMS-1564	SM3, La Chacra	7930	30	8972	8590	8951	8586	8672	Charcoal	155
Poz-38865	SM3, La Chacra	7860	50	8847	8435	8682	8426	8557	Shell	236
Poz-38866	SM3, La Chacra	7790	80	8760	8383	8748	8456	8590	Charcoal	240
Poz-38853	SM2, San Francisco	4950	40	5736	5585			5638	Shell	85
Poz-38850	SM2, San Francisco	4770	60	5588	5320			5465	Charcoal	155
Poz-38851	SM2, San Francisco	5380	40	6272	5994			6117	Shell	205
Poz-38852	SM2, San Francisco	5500	40	6393	6128			6253	Charcoal	205
Poz-34228	SM1, Isla del Tesoro	345	25	452	304			391	Bone	15
Poz-34229	SM1, Isla del Tesoro	3895	35	4411	4153	4414	4155	4292	Shell	35
Poz-34230	SM1, Isla del Tesoro	4945	35	5722	5586	5721	5585	5629	Shell	45
Poz-28854	SM1, Isla del Tesoro	3830	50	4405	3986			4172	Shell	48
Poz-28855	SM1, Isla del Tesoro	4415	35	5210	4849			4937	Charcoal	77
Poz-22902	SM1, Isla del Tesoro	5520	40	6395	6190			6280	Charcoal	115
Poz-34231	SM1, Isla del Tesoro	5520	40	6395	6190	6295	6190	6234	Shell	115
Poz-24633	SM1, Isla del Tesoro	5360	40	6266	5950			6096	Shell	115
D-AMS-1740	SM1, Isla del Tesoro	5502	30	6310	6190	6309	6207	6275	Charcoal	117
PSUAMS-4658	SM1, Isla del Tesoro	5565	20	6398	6281	6393	6278	6305	Charcoal	120
Poz-24634	SM1, Isla del Tesoro	5505	35	6388	6184			6262	Charcoal	120
Poz-34232	SM1, Isla del Tesoro	5460	40	6304	6021	6302	6205	6254	Shell	125
Poz-28856	SM1, Isla del Tesoro	4480	40	5284	4872			5047	Charcoal	140
Poz-28850	SM1, Isla del Tesoro	4495	35	5288	4886			5109	Shell	140
Poz-36136	SM1, Isla del Tesoro	5800	35	6657	6453	6660	6461	6563	Charcoal	160
D-AMS 032885	SM1, Isla del Tesoro	7271	40	8162	7966	8151	7956	8019	bulk organic	205
D-AMS 032884	SM1, Isla del Tesoro	7447	37	8348	8065	8347	8174	8274	bulk organic	225
Poz-34301	SM1, Isla del Tesoro	9270	60	10556	10248	10573	10259	10434	bulk organic	235
Poz-36135	SM1, Isla del Tesoro	9420	50	10743	10433	10715	10416	10574	Charcoal	245
BE-4254.1.1	FIN8 182-187	8681	22	9671	9537			9582	Charcoal	185
BE-4253.1.1	FIN5 92-97	6963	25	7828	7680			7745	Shell	95
BE-4250.1.1	FIN3 95-100	8572	48	9560	9438			9514	Charcoal	97
BE-4257.1.1	FIN15 145-150	7997	25	8985	8649			8834	Shell	147
BE-4256.1.1	FIN14 85-90	6552	25	7486	7325			7429	Shell	87
BE-4255.1.1	FIN12 90-95	4092	24	4785	4424			4522	Shell	92
BE-8256.1.1	BANR17-UE1-5	3017	21	3236	3005	3319	3007	3144	Charcoal	25
BE-8257.1.1	BANR17-UE1-20	4324	22	4959	4743	4957	4728	4848	Charcoal	55
BE-8258.1.1	BANR17-UE1-31	4491	23	5285	4888	5288	4965	5172	Charcoal	85
BE-8259.1.1	BANR17-UE1-38	9138	24	10367	10195	10284	10196	10239	Charcoal	105
	BANR17-UE1-53					10546	10231	10372		125
D-AMS 029221	BANR17-UE1-57	9346	41	10653	10298	10664	10380	10521	Charcoal	145
	BANR17-UE1-64					13918	10307	10883		155
BE-7663.1.1	575 270-280	8849	50	10155	9632			9848	Charcoal	275
BE-7671.1.1	574 80-90	5516	62	6409	6025			6273	Charcoal	85
BE-7667.1.1	570 60-70	6046	48	6984	6695			6844	Charcoal	65
BE-7675.1.1	569 40-50	407	19	496	328			453	Charcoal	45
BE-7672.1.1	569 130-140	5759	53	6659	6399			6512	Charcoal	135
BE-7664.1.1	536 110-120	4994	37	5857	5595			5678	Charcoal	115
BE-7668.2.1	535 190-200	6069	51	7145	6730			6871	Charcoal	195
BE-7673.2.1	530 160-170	2397	20	2465	2327			2362	Charcoal	165
BE-7661.1.1	527 200-210	5337	281	6730	5472			6076	Charcoal	205
BE-7662.1.1	526 250-260	6217	40	7239	6945			7076	Charcoal	255
BE-7674.1.1	521 90-100	2346	89	2701	2094			2331	Charcoal	95
BE-7666.1.1	519 100	5647	22	6443	6308			6368	Charcoal	100
BE-6164.1.1	502 65	4365	110	5295	4583			4927	Charcoal	65
BE-6166.1.1	493 155	8920	49	10186	9766			10006	char/sed	155
BE-6153.1.1	490 95-100	3796	33	4237	3985			4115	Shell	97
BE-6167.1.1	434 270	7811	57	8696	8413			8543	char/sed	270
BE-6163.1.1	433 175	7058	50	7954	7718			7852	charcoal	175
BE-6168.1.1	430 235	6397	130	7552	6951			7269	char/sed	235
BE-6158.1.1	429 220	8028	24	8999	8717			8870	char/sed	220
BE-6157.1.1	423 170	4365	21	4965	4840			4877	char/sed	170
BE-6159.1.1	421 170	5875	127	6953	6322			6646	char/sed	170

Stratigraphically ordered dated depths have been modelled using a Bayesian age–depth model (the P_Sequence of OxCal v.4.3). The modelled ages for samples BANR17-UE1-53 and BANR17-UE1-64 have been calculated using the Date command. Code is available in the Supplementary Information.

Extended Data Table 2 | Length and thickness range and average size of scalloped-sphere phytoliths identified in this study

Site	Date cal. yr. B.P.	Length (µm)	Thickness (µm)	Domesticated	Depth
Isla del Tesoro	7839-7020	80,361	59,804	Yes	170-180 cm
575	10155-9632	82,524	58,318	Yes	270-280 cm
		72,804	65,178	Yes	
		72,775	54,214	Yes	
		77,884	55,212	Yes	
519	6443-6380	82,418	65,752	Yes	100 cm
		75,938	54,452	Yes	
		75,296	57,433	Yes	
		81,19	55,162	Yes	
FIN-12	4785-4424	80,653	60,988	Yes	90-95 cm
		61,005	45,635	No	
		77,871	67,843	Yes	
		83,697	59,724	Yes	
La Chacra	6707-5826	78,895	52,993	Yes	80-84 cm
	7350-6945	75,789	56,754	Yes	108-113 cm
	8530-7965	63,119	41,238	No	140-145 cm
	9212-8665	75,419	65,484	Yes	160-165 cm
	3319-3007	74.28	52.33	Yes	25 cm
		72.7	57.57	Yes	
		78.49	50.79	Yes	
		81.34	48.89	Yes	
		70.1	45.38	Yes	
		78.12	64.79	Yes	
		61.29	51	No	
		69.22	56.6	No	
		61.78	46.56	No	
		59.96	49.13	No	
		68.67	54.95	No	
		57.22	40.23	No	
	4957-4728	70.53	49.5	No	55 cm
		63.21	49.51	No	
		70.59	44.29	No	
		81.74	66.85	Yes	
		72.76	58.61	Yes	
		74.27	57.12	Yes	
		74.53	45.13	Yes	
		58.98	44.15	No	
	5288-4965	81.95	56.8	Yes	85 cm
		68.1	59.06	No	
		66.84	55.37	No	
		76.91	49.26	Yes	
		74.69	47.15	Yes	
		84.72	60.19	Yes	
		78,173	66,108	Yes	105 cm
		84,996	67,836	Yes	
		72,601	54,422	Yes	
		64,201	50,158	No	
		84,858	64,479	Yes	
		68.44	57.37	No	
		81.82	60.62	Yes	
	10284-10196	86.81	63.17	Yes	105 cm
Isla Manechi	10664-10380	60.46	51.16	No	145 cm

Based on a previous study⁴⁹, we consider scalloped spheres longer than 72 µm or thicker than 59 µm as coming from domestic varieties of *Cucurbita* sp.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☒ ☐ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Forest Islands and per-Columbian earthworks have been mapped using ArcGIS pro 2.4

Data analysis

Spatial data have been analysed using ArcGIS pro 2.4; radiocarbon ages have been calibrated and modelled using OxCal 4.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data needed to evaluate the conclusions in the paper are included in the paper or in the Extended Data. Code used for 14C calibration in OxCal is available in Supplementary Information.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Data are quantitative. The study includes mapping of landscape elements based on visual analysis of remote sensing data; sampling of forest islands; 14C of the FIs' samples; phytolith extraction and visual (microscope) analysis (counting of individual phytoliths) of FIs' samples
Research sample	Samples are sediments cored on forest islands using an auger. In four cases (SM1, SM3, SM4 and Manechi) samples have been taken from profiles exposed during archaeological excavations. Samples are representative for the whole region because the four areas we surveyed cover all the different eco-regions identified in the Llanos de Moxos.
Sampling strategy	For archaeological excavations, samples have been taken from stratigraphic profiles at different depths. For the rest of forest islands, samples have been analysed from the lowest (i.e. oldest) datable level. FIs have been chosen for sampling based on their location. The two main criteria have been assuring the representativeness of the total population (by sampling 4 different regions) and accessibility. The amount of FIs sampled was not pre-determined. We sampled the maximum number of sites we could sample within our logistic constraints.
Data collection	Samples have been taken with an auger. After extraction from the subsoil the excess of material has been cut off with a knife and only the inner, uncontaminated part of the extracted samples have been stored in plastic bags. These have been air-dried in Bolivia before being shipped. Charcoal fragments for 14C have been collected in situ, enveloped in aluminium foil and stored in plastic bags. Field observations have been wrote down on a notebook. The researchers were aware of the study hypothesis at the time of sampling.
Timing and spatial scale	Sampling has been done in different field seasons for different sites. The totality of the samples have been taken in 2012, 2013, 2014, 2016 and 2017. Samples have been taken in the Beni department, Bolivia.
Data exclusions	We decided to discard the identification of maize based of statistical analysis of cross-shaped phytoliths in order to rely only on the presence of wavy top rondel phytoliths. This exclusion was not pre-established. Maize can be identified using a discriminant function on non diagnostic phytoliths (cross shaped) or by identifying diagnostic phytoliths (wavy rondel), as we did for the rest of cultivars. Diagnostic phytoliths are a direct evidence far more reliable than the discriminant function. The discriminant function on cross shaped phytoliths indicate presence of maize in samples dated ca. 10.000 BP. It is almost impossible that maize was present in Bolivia 10k yrs ago, as this would precede the time of its domestication in Mexico. It could be that in this particular context (Bolivian Amazon in the early Holocene) some other plant produced maize-like cross-shaped phytoliths which affected our discriminant analysis. In the early contexts we did not find the diagnostic wavy-top rondels phytoliths derived from the maize glumes which were found in the other later samples. This need more research and we plan to further investigate this issue.
Reproducibility	The experiments consisted in counting a standard number (200) of diagnostic phytoliths. This number is considered sufficient to be representative of the sample, therefore it is not standard practice to repeat the counting.
Randomization	Sampling was not completely random because we choose the forest islands also based on their accessibility and ownership of the land. However, none of these criteria affect the representativeness of our sample.
Blinding	Sampling was not blind because we sampled soil and subsoil, so we knew the origin of each sample. In the lab samples where coded with numbers. Sample extraction and phytolith counting was blind because the origin of the sample was unknown during these steps.
Did the study involve field work?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

Field work, collection and transport

Field conditions	Sampling has always being performed during the dry season, between July and September. No forest islands has been sampled while raining.
Location	Fieldwork took place in the Beni department, Bolivia. The area surveyed is enclosed in a square area: up right corner lat -13° Lon -67°; down left corner Lat -15.5°; Lon -63.5°. Average elevation 180 m a s l. All sampling was performed on land.
Access and import/export	Field sites have been accessed with the permission of the land owner. Authorizations have been obtained by the Bolivian Ministry of Cultures and Tourism (UDAM 017/2012, UDAM 027/2013, 019/2014, UDAM 006/2015, UDAM 071/2017) and by the Beni Autonomous Government (08/08/2013, DDT 64-A/2014 and DDT 138/2017).

Disturbance

Forest islands were accessed by walking. Cores were taken with a manual auger with a 5 cm diameter, archaeological excavation were performed with manual tools, minimizing noises and impact on local fauna. The excavation pits were refilled with the excavated sediments in order to restore the aspect of the sites previous to the excavation.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Retinal innervation tunes circuits that drive nonphotic entrainment to food

<https://doi.org/10.1038/s41586-020-2204-1>

Received: 23 July 2019

Accepted: 21 February 2020

Published online: 22 April 2020

 Check for updates

Diego Carlos Fernandez^{1✉}, Ruchi Komal¹, Jennifer Langel¹, Jun Ma¹, Phan Q. Duy^{1,3}, Mario A. Penzo¹, Haiqing Zhao² & Samer Hattar^{1✉}

Daily changes in light and food availability are major time cues that influence circadian timing¹. However, little is known about the circuits that integrate these time cues to drive a coherent circadian output^{1–3}. Here we investigate whether retinal inputs modulate entrainment to nonphotic cues such as time-restricted feeding. Photic information is relayed to the suprachiasmatic nucleus (SCN)—the central circadian pacemaker—and the intergeniculate leaflet (IGL) through intrinsically photosensitive retinal ganglion cells (ipRGCs)⁴. We show that adult mice that lack ipRGCs from the early postnatal stages have impaired entrainment to time-restricted feeding, whereas ablation of ipRGCs at later stages had no effect. Innervation of ipRGCs at early postnatal stages influences IGL neurons that express neuropeptide Y (NPY) (hereafter, IGL^{NPY} neurons), guiding the assembly of a functional IGL^{NPY}–SCN circuit. Moreover, silencing IGL^{NPY} neurons in adult mice mimicked the deficits that were induced by ablation of ipRGCs in the early postnatal stages, and acute inhibition of IGL^{NPY} terminals in the SCN decreased food-anticipatory activity. Thus, innervation of ipRGCs in the early postnatal period tunes the IGL^{NPY}–SCN circuit to allow entrainment to time-restricted feeding.

The circadian system contains the SCN, the central circadian pacemaker, that orchestrates rhythmic functions of peripheral clocks located throughout the body¹. This system integrates multiple time cues from sensory, as well as circadian and metabolic, systems to generate a coherent perception of the environment^{2,3}. At present, little is known about the brain circuits and mechanisms that integrate different time cues to drive a coordinated circadian output.

In mammals, light is transmitted to circadian centres through a subpopulation of retinal ganglion cells⁴ that are intrinsically photosensitive, owing to their expression of the photopigment melanopsin (encoded by *OPN4*)^{5,6} to drive circadian photoentrainment. As a central pacemaker, the SCN receives dense axonal projections from multiple areas of the brain—particularly the thalamic IGL, which is thought to be involved in the circadian entrainment to nonphotic cues (hereafter, nonphotic entrainment)^{7–10}. Here we show that retinal input affects circadian circuits that control nonphotic entrainment.

Ablation of ipRGCs attenuates timed feeding

We assessed the effect of ablating retinal input on the circuits that control entrainment to nonphotic time cues. We used a mouse line that removes ipRGCs during development up to early postnatal stages through the expression of subunit A of diphtheria toxin (*Opn4^{DTA}* mice)¹¹. Nonphotic entrainment was evaluated by limiting the food access to a 7-h period (Fig. 1a), in what is known as time-restricted feeding (TRF)¹². We opted to keep mice under constant darkness; thus, the time-restricted access to food constituted the only recurrent time cue

for the mice. Both female and male control (wild-type) and *Opn4^{DTA}* mice with ad libitum access to food (hereafter, free-running conditions) showed robust rhythmic patterns of feeding that closely overlapped with their patterns of locomotor activity (Extended Data Fig. 1a, b), which confirms that early ablation of ipRGCs has no effect on locomotor activity and rhythmic feeding pattern in adult mice.

Under TRF, control mice displayed a robust and sustained food-anticipatory activity^{13,14} (Fig. 1b–d, Extended Data Fig. 1c). *Opn4^{DTA}* mice showed deficits in nonphotic entrainment to TRF (Fig. 1b–d, Extended Data Fig. 1d), as reduced food-anticipatory activity was observed throughout the restriction paradigm (Fig. 1e, f). A graded-score analysis system (Extended Data Fig. 1e; see Methods for a full description of the system) similarly showed significant deficits in circadian anticipation to TRF in mice that lack ipRGC innervation (Fig. 1g).

We next evaluated the hormones involved in the control of feeding. Levels of insulin, leptin and total ghrelin—as well as glucose—were similar in control and *Opn4^{DTA}* mice under free-running conditions (Extended Data Table 1). Under TRF, control and *Opn4^{DTA}* mice had similar levels of glucose and anorexigenic hormones, leptin and insulin (Fig. 1h, i, Extended Data Table 1). In addition, control and *Opn4^{DTA}* mice consumed similar amounts of food, and their feeding patterns, body weight and body composition were indistinguishable (Extended Data Fig. 1f–j). Together, these results indicate that the behavioural alterations that we observed in mice that lack ipRGCs are not caused by changes in food intake or caloric restriction.

The levels of total ghrelin—an orexigenic hormone known for its stimulatory effects on food intake^{15,16}—were increased in anticipation

¹National Institute of Mental Health (NIMH), National Institutes of Health (NIH), Bethesda, MD, USA. ²Department of Biology, Johns Hopkins University, Baltimore, MD, USA. ³Present address: MSTP, Yale University, New Haven, CT, USA. ✉e-mail: diego.fernandez@nih.gov; samer.hattar@nih.gov

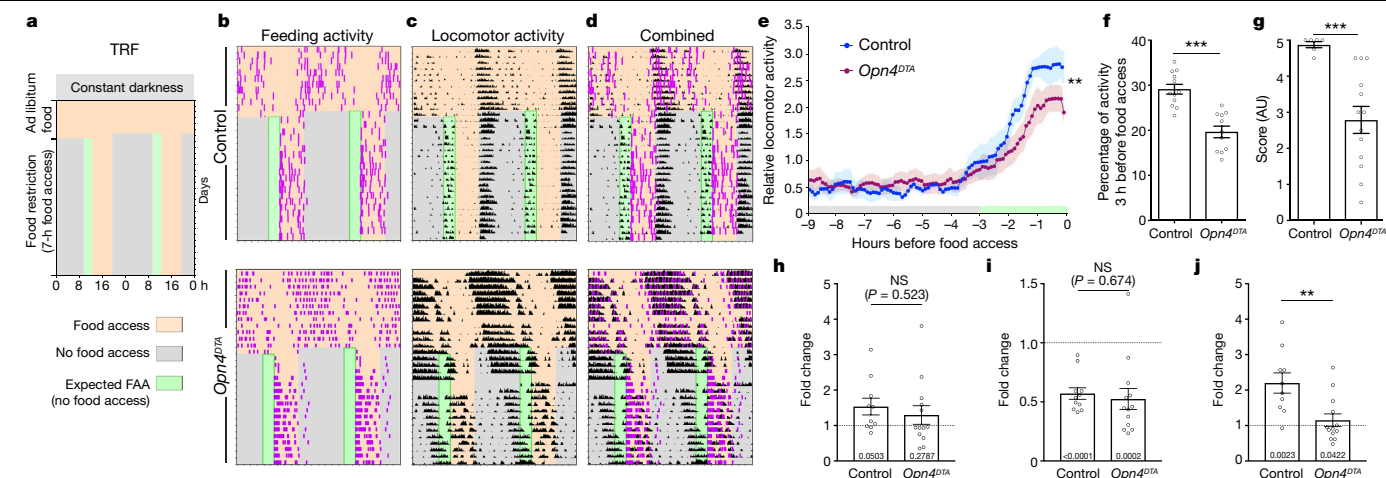


Fig. 1 | Early ablation of ipRGCs affects circadian entrainment to TRF. **a**, Schematic illustrating the TRF paradigm. FAA, food-anticipatory activity. **b–d**, Representative actograms for feeding (**b**), locomotor activity (**c**) and combined actograms (**d**), measured in three-month-old control and *Opn4*^{DTA} mice. **e, f**, The locomotor activity before food access (**e**), and relative food-anticipatory activity (**f**) were measured as described in Methods. Data are mean \pm s.e.m. ($n=12$ control mice, 13 *Opn4*^{DTA} mice), *** $P=0.0001$, ** $P=0.0042$, two-tailed Student's *t*-test. **g**, A score analysis was performed for all actograms obtained from control and *Opn4*^{DTA} mice as described in Methods. Data are

mean \pm s.e.m. ($n=12$ control mice, 13 *Opn4*^{DTA} mice), *** $P=0.003$, Student's *t* non-parametric (Mann–Whitney) test, two-tailed. AU, arbitrary units. **h–j**, Hormonal signals were measured in control and *Opn4*^{DTA} mice. The levels of leptin (**h**), insulin (**i**), and total ghrelin (**j**) were measured. Data are expressed as the level of the hormone during restricted access to food relative to free-running conditions. Data are mean \pm s.e.m. ($n=10$ control mice, 13 *Opn4*^{DTA} mice), ** $P=0.0038$, two-tailed Student's *t*-test. NS, not significant. In addition, the statistical analysis versus a hypothetical value = 1 (dotted lines) was performed, and is shown per column; by one-sample *t*-test.

of food in control, but not *Opn4*^{DTA} mice (Fig. 1j, Extended Data Table 1). The differences in total ghrelin levels between groups were observed as early as day 7 after TRF and persisted throughout the experiments (Extended Data Fig. 1k). These results indicate that the lack of anticipatory ghrelin responses in *Opn4*^{DTA} mice correlate with the impaired anticipatory activity to timed feeding.

ipRGCs influence the IGL–SCN circuit

Several brain and peripheral areas exhibit changes in activity in response to TRF¹⁷. The reduced nonphotic entrainment displayed by *Opn4*^{DTA} mice suggests the involvement of a target of ipRGCs. The IGL receives dense innervation from ipRGCs and is implicated in driving photic and nonphotic signals to modulate circadian processes^{7,18}. We found that mice exposed to TRF showed a substantial induction of the immediate-early gene *Fos* (also known as *c-Fos*) in neurons of the IGL (Fig. 2a–c). However, *Opn4*^{DTA} mice exposed to the same paradigm showed a reduced *FOS* induction in the IGL (Fig. 2b, c). Importantly, in control and *Opn4*^{DTA} mice, *FOS* induction was not observed in the hypothalamic arcuate nucleus—an area known to be involved in the homeostatic control of hunger and food intake¹⁹ (Extended Data Fig. 2a–c). These results implicate the IGL as a brain region involved in circadian entrainment to TRF.

The IGL contains NPY-expressing neurons that project to the SCN^{7–9}. To better characterize the innervation pattern of IGL^{NPY} neurons, we injected a Cre-dependent adeno-associated virus (AAV), AAV-DIO-tdTomato, into the IGL of *Npy*^{cre/+} mice. We found that IGL^{NPY} neurons innervate both the ipsi- and contralateral SCN and, to a lesser extent, send unilateral projections to other regions of the brain (Extended data Fig. 2d–h). Early ablation of ipRGCs causes a significant reduction in the NPY immunoreactivity in IGL neurons (Fig. 2d, e), whereas the number of NPY⁺ somas and DAPI⁺ nuclei were unaffected (Fig. 2f, g). Consistent with the reduction in NPY levels in the IGL, we also found a significant reduction in NPY⁺ reactivity in the SCN of *Opn4*^{DTA} mice (Fig. 2h–j, Extended data Fig. 2i, j, Supplementary Videos 1, 2). However, when we correlated the NPY levels from the SCN and the IGL, we found that the NPY staining in the IGL covers a larger percentage of the leaflet compared to the SCN nuclei in *Opn4*^{DTA} mice (Extended data Fig. 2j), suggesting that NPY axonal

transport could also be affected. Structures that are not innervated by ipRGCs and that express high levels of NPY showed normal patterns of NPY immunostaining (Extended data Fig. 2k).

Time window for IGL–SCN circuit assembly

We next evaluated whether ablation of ipRGCs at adult stages causes similar alterations to responses to TRF. We used a mouse line that expresses an attenuated form of the diphtheria toxin (*atnD*^{DTA}, also known as *aDTA*) controlled by the melanopsin promoter (hereafter, *Opn4*^{atnD} mice), inducing substantial ablation of ipRGCs that project to the SCN and IGL by six months of age⁴. Eliminating the innervation by ipRGCs in adult mice had no significant effect on NPY levels in fibres innervating the SCN (Fig. 3a, b). In addition, adult *Opn4*^{atnD} mice exposed to the TRF protocol showed robust food-anticipatory activity, comparable to age-matched controls (Fig. 3c–e, Extended Data Fig. 3a, b). These results indicate that innervation by ipRGCs has an important role in circuits that control entrainment to TRF, specifically during early postnatal stages.

To determine the critical period for the influence of innervation by ipRGCs on the assembly of the IGL^{NPY}–SCN circuit and nonphotic entrainment, we enucleated wild-type mice at different postnatal stages. We found that the IGL^{NPY}–SCN circuit was disrupted in adult mice that were enucleated at postnatal day (P)0 to up to P40, an early adulthood stage. However, enucleation of mice at P90, a mature adulthood stage had no significant effect (Fig. 3f, g, Extended Data Fig. 3c, Supplementary Videos 3–6). Concordantly, wild-type mice enucleated at P0 and P40 showed significant deficits in the entrainment to TRF, whereas mice enucleated at P90 displayed robust food-anticipatory responses (Fig. 3h–l, Extended Data Fig. 3d–g). These results indicate that there is a critical time window for the innervation by ipRGCs to influence the assembly of the IGL^{NPY}–SCN circuit and nonphotic entrainment.

ipRGC–SCN axons regulate the IGL–SCN circuit

SCN and IGL receive dense input from ipRGCs^{8,9}, which suggest that ipRGCs could directly affect IGL^{NPY} neurons, their axonal projections

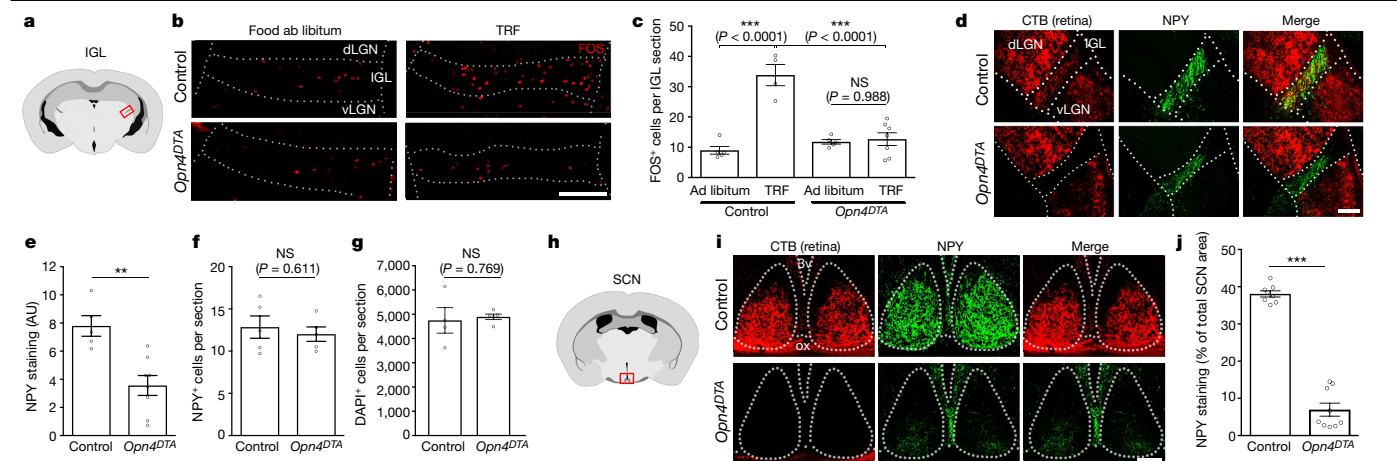


Fig. 2 | Early ablation of ipRGCs alters connectivity between IGL^{NPY} and SCN. **a**, Schematic brain section, highlighting the location of the IGL.

b, c, Representative images showing FOS induction in IGL neurons in response to ad libitum food or TRF in three-month-old control and *Opn4^{DTA}* mice (**b**), quantified in **c**. Data are mean \pm s.e.m. ($n = 4$ mice for each genotype), *** $P < 0.001$, two-tailed Tukey's test. **d–g**, Morphological characterization of retinal innervation (CTB, red) and NPY staining (green) in the IGL in three-month-old control and *Opn4^{DTA}* mice. Representative coronal sections are shown (**d**). The levels of NPY staining (**e**), number of NPY⁺ somas (**f**) and

DAPI⁺ nuclei (**g**) were quantified. Data are mean \pm s.e.m. ($n = 5$ control mice, 8 *Opn4^{DTA}* mice), ** $P = 0.0022$, two-tailed Student's *t*-test. **h**, Schematic brain section highlighting the location of the SCN. **i, j**, Retinal input to the SCN (CTB, red) and NPY from IGL (NPY, green). Representative images are shown (**i**). NPY staining in the SCN was quantified (**j**). Data are mean \pm s.e.m. ($n = 8$ mice for each genotype), *** $P < 0.001$, two-tailed Student's *t*-test. dLGN, dorsal lateral geniculate; vLGN, ventral lateral geniculate; 3v, third ventricle; ox, optic chiasm. Scale bars, 100 μ m (**i**), 200 μ m (**b, d**).

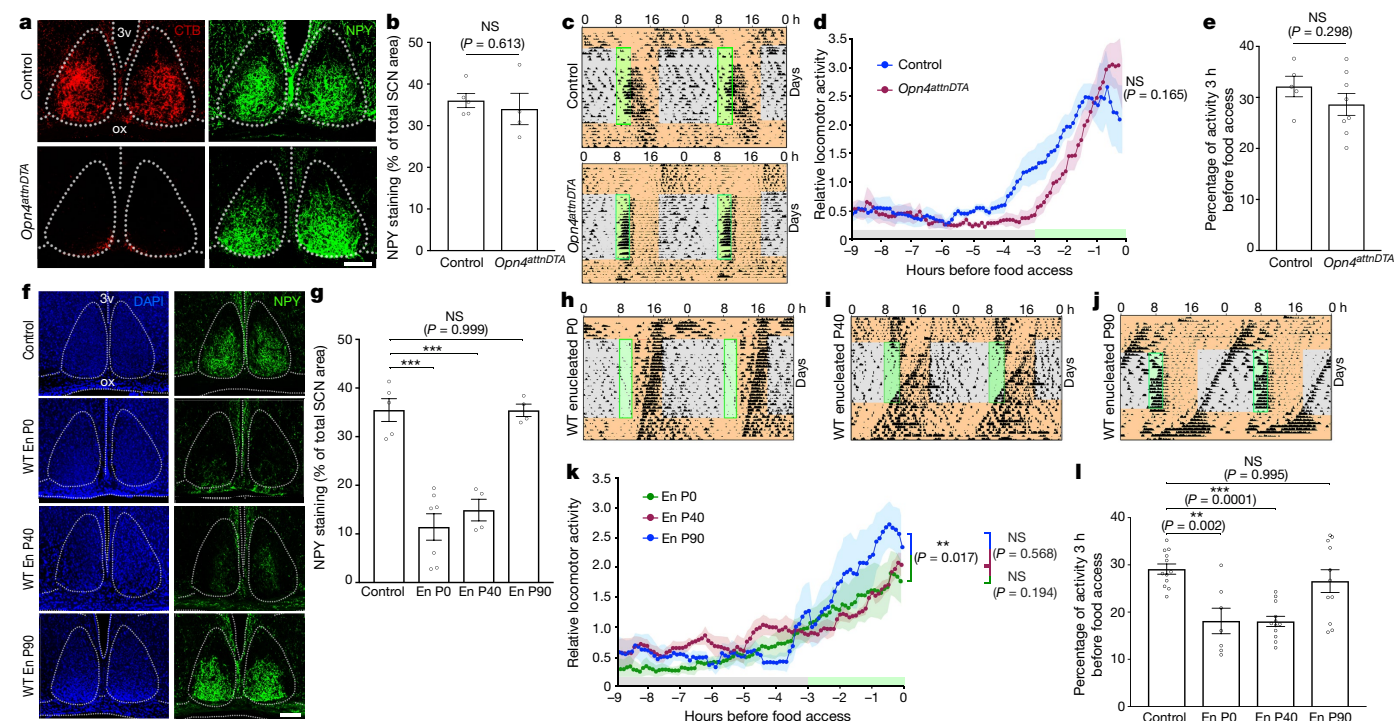


Fig. 3 | Critical time window for assembly of the IGL^{NPY}–SCN circuit.

a, b, Retinal (CTB, red) and IGL (NPY, green) innervation to SCN in nine-month-old control and *Opn4^{atnDTA}* mice. Representative SCN sections are shown (**a**). NPY staining was analysed (**b**). Data are mean \pm s.e.m. ($n = 5$ mice for each genotype), two-tailed Student's *t*-test. **c–e**, Locomotor activity was measured in control and *Opn4^{atnDTA}* mice exposed to TRF. Representative actograms are shown (**c**). The locomotor activity before food access (**d**) and the food-anticipatory activity (**e**) were measured. Data are mean \pm s.e.m. ($n = 5$ control mice, 8 *Opn4^{atnDTA}* mice), two-tailed Student's *t*-test. **f, g**, IGL

(NPY, green) innervation to SCN in 3–5-month-old in wild-type (WT) control and enucleated (En) mice. Representative sections are shown (**f**). NPY staining in the SCN was analysed (**g**). Data are mean \pm s.e.m. ($n = 5$ control, 7 En P0, 4 En P40 and 5 En P90 mice); *** $P < 0.001$, two-tailed Tukey's test. **h–j**, Wild-type mice, enucleated at different stages, were exposed to TRF. Representative actograms are shown (**h–j**). The locomotor activity before food access (**k**) and the food-anticipatory activity (**l**) were measured in intact and enucleated mice. Data are mean \pm s.e.m. ($n = 11$ control, 7 En P0, 12 En P40 and 11 En P90 mice), two-tailed Tukey's test. Scale bars, 100 μ m (**a, f**).

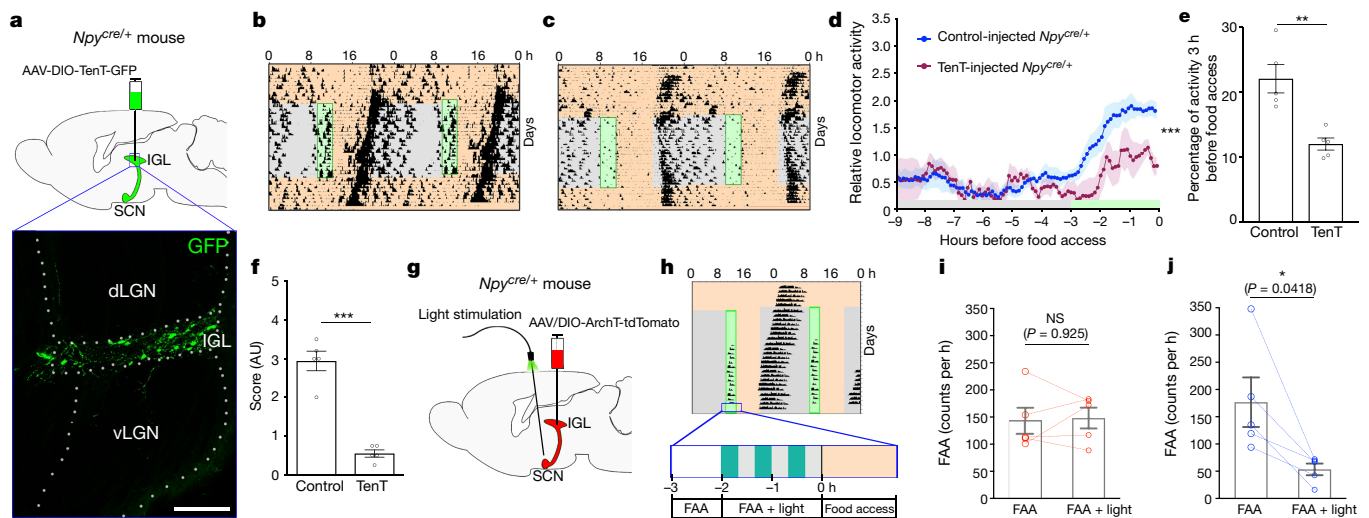


Fig. 4 | NPY signalling in the IGL-SCN circuit regulates entrainment to TRF.

a, A Cre-dependent AAV encoding TenT-GFP was bilaterally injected into the IGL of three-month-old *Npy^{cre/+}* mice. Top, schematic. Bottom, injection site. **b, c**, Representative actograms obtained from *Npy^{cre/+}* mice exposed to TRF that were injected with a control AAV (**b**) and AAV encoding TenT (**c**). **d–f**, Locomotor activity before food access (**d**) and food-anticipatory activity (**e**) were measured for control and *Npy^{cre/+}* mice. Data are mean \pm s.e.m. ($n = 5$ mice for each genotype), *** $P < 0.001$, ** $P = 0.0029$, two-tailed Student's *t*-test. A score analysis was performed for all actograms obtained (**f**); data are mean \pm s.e.m. ($n = 5$ mice for each genotype); *** $P = 0.0079$, Student's

t non-parametric (Mann-Whitney) test, two-tailed. **g–h**, Optogenetic silencing of the IGL^{NPY}-SCN circuit during food anticipatory activity. Schematic showing the AAV injections and cannula implantation (**g**). Schematic showing the optical stimulation protocol (**h**). **i, j**, Locomotor activity was measured for 3 h (in 5-min bins) before food access in *Npy^{cre/+}*-sham (**i**) and *Npy^{cre/+}* ArchT (**j**) mice exposed to TRF and optogenetic stimulation. Data are expressed as activity (total beam break counts per hour) measured during food anticipatory activity, with or without optical stimulation. Data are mean \pm s.e.m. ($n = 5$ mice for each condition), * $P = 0.0418$, two-tailed paired Student's *t*-test. Scale bar, 200 μ m (**a**).

at the SCN level or a combination of both. To test these possibilities, we used a mouse line (*Opn4^{cre/+} Brn3b^{DTA/+}*) (*Brn3b* is also known as *Pou4f2*) in which *Brn3b⁺* ipRGCs that innervate the SCN survive (Extended Data Fig. 4a), whereas *Brn3b⁺* ipRGCs that mostly project to non-SCN regions (including the IGL) are ablated during early postnatal stages^{20,21}. In *Opn4^{cre/+} Brn3b^{DTA/+}* mice, NPY immunostaining was unaffected in both the IGL and SCN (Extended Data Fig. 4a–c), and these mice showed sustained food-anticipatory activity to TRF (Extended Data Fig. 4d–h). Thus, innervation by ipRGCs to the SCN is sufficient for tuning the functional assembly of the IGL^{NPY}-SCN circuit.

ipRGCs establish NPY levels in IGL

The reduced NPY levels in fibres that innervate the SCN suggest that there is either a marked depletion of the neuropeptide or a lack of axonal innervation from the IGL. To test this, we enucleated *Npy^{cre/+}* mice at P0, and three months later mice were injected in the IGL with a Cre-dependent AAV to trace NPY⁺ projections and their synaptic terminals (Extended Data Fig. 5a, b). We found that the innervation pattern and the density of synaptic terminals at the SCN level were not affected in adult *Npy^{cre/+}* mice that were enucleated at P0 (Extended Data Fig. 5c, d), demonstrating that early ablation of ipRGCs affects the level of NPY, but not IGL^{NPY} axonal projections.

IGL^{NPY} neurons affect entrainment to TRF

Adult mice that lack NPY (NPY-knockout mice) showed reduced entrainment to TRF (Extended Data Fig. 6a–d). In addition, NPY-knockout mice showed a reduction in the amount of food they consumed during the TRF paradigm (Extended Data Fig. 6e), reflecting the critical role that NPY signalling has in overall feeding behaviour²².

To silence the synaptic release of IGL^{NPY} neurons, a Cre-dependent AAV encoding the light chain subunit of tetanus toxin (TenT) fused to

green fluorescent protein (GFP) was injected into adult *Npy^{cre/+}* mice (Fig. 4a). *Npy^{cre/+}* mice injected with an AAV encoding DIO-GFP were used as a control group. Injection sites were confirmed post hoc by assessing GFP expression (Fig. 4a). Four weeks after the AAV injections, mice were exposed to the TRF paradigm. The inhibition of synaptic release specifically in IGL^{NPY} neurons abolished food-anticipatory activity (Fig. 4b–f, Extended Data Fig. 6f, g). It is important to note that, similar to NPY-knockout mice, the amount of food consumed by *Npy^{cre/+}* mice was reduced when housed with ad libitum access to food (Extended Data Fig. 6e). However, this reduction was exacerbated during the time-restricted access to food in *Npy^{cre/+}* mice injected with the AAV encoding TenT (Extended Data Fig. 6e).

To specifically manipulate the IGL^{NPY}-SCN circuit, we used a virally delivered optogenetic strategy to transiently silence IGL^{NPY} projections that innervate the SCN. A Cre-dependent AAV encoding archaerhodopsin TP009 fused with tdTomato (DIO-ArchT-tdTomato) was bilaterally injected in adult *Npy^{cre/+}* mice (hereafter, *Npy^{cre/+}* ArchT mice), and optical fibres were implanted just above the SCN (Fig. 4g). As control group, *Npy^{cre/+}* mice were injected with an AAV encoding DIO-tdTomato (hereafter, *Npy^{cre/+}* sham mice) (Extended Data Fig. 7a). Mice were allowed to recover for two weeks, and then placed under TRF. Both *Npy^{cre/+}* sham mice and *Npy^{cre/+}* ArchT mice showed similar food-anticipatory activity and food consumption (Extended Data Fig. 7b–d). Next, neural silencing was optogenetically induced in the IGL^{NPY}-SCN circuit starting 2 h before food delivery by applying 3 pulses of 20 min of light, with 20-min intervals (Fig. 4h). Locomotor activity was measured starting 3 h before food delivery, and the total activity was recorded before and after neuronal silencing, and compared in both groups of mice. Control *Npy^{cre/+}* sham mice displayed no significant changes in the locomotor activity (Fig. 4i, Extended Data Fig. 7e, g, h). By contrast, *Npy^{cre/+}* ArchT mice displayed reduced exploratory activity during the optical silencing of NPY^{ArchT}-positive fibres (Fig. 4j, Extended Data Fig. 7f–h), without affecting subsequent feeding behaviour (Extended Data Fig. 7c, d).

These results demonstrate that the signalling of IGL^{NPY} neurons projecting specifically to SCN is required for entrainment to TRF.

Discussion

Here we reveal that early innervation by ipRGCs affects the assembly of a functional IGL^{NPY}–SCN circuit. Moreover, the correct assembly of this circuit is necessary for the circadian anticipatory responses that are associated with TRF in adult mice.

The integration of environmental cues occurs widely in the nervous system. It has been extensively documented that early ablation of retinal input to the superior colliculus—a major node of multisensory integration²³—leads to an extensive rewiring of its sensory inputs²⁴, causing the strengthening of responses to nonvisual stimuli^{25,26}. On the basis of these results, our original expectation was that ablating the innervation by ipRGCs to brain centres would cause a strengthening of circadian entrainment to nonphotic modalities. Our results show a weakening of nonphotic entrainment in mice that lack early retinal innervation to brain targets. Therefore, we propose that the retina–IGL^{NPY}–SCN circuits do not follow the conventional model of early plasticity that has been described for the image-forming visual system.

Feeding behaviour is the result of the integration of multiple responses to food, including energy homeostasis²⁷, hedonic reward²⁸ and memory traces²⁹, as well as anticipation to timed food availability driven by a food-entrainable oscillator^{30,31}. Our results suggest that IGL^{NPY} neurons require retinal innervation to the SCN during early development for normal circuit assembly, and act as a node of connection between the food-entrainable-oscillator network and the central pacemaker of the SCN in adult mice (Extended Data Fig. 8, Supplementary Discussion).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2204-1>.

1. Golombek, D. A. & Rosenstein, R. E. Physiology of circadian entrainment. *Physiol. Rev.* **90**, 1063–1102 (2010).
2. Cappe, C., Rouiller, E. M. & Barone, P. In *The Neural Bases of Multisensory Processes* (eds Murray, M. M. & Wallace, M. T.) Chapter 2 (CRC, 2012).
3. Mahoney, J. R. et al. Keeping in touch with the visual system: spatial alignment and multisensory integration of visual-somatosensory inputs. *Front. Psychol.* **6**, 1068 (2015).
4. Güler, A. D. et al. Melanopsin cells are the principal conduits for rod-cone input to non-image-forming vision. *Nature* **453**, 102–105 (2008).

5. Hattar, S., Liao, H. W., Takao, M., Berson, D. M. & Yau, K. W. Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science* **295**, 1065–1070 (2002).
6. Berson, D. M., Dunn, F. A. & Takao, M. Phototransduction by retinal ganglion cells that set the circadian clock. *Science* **295**, 1070–1073 (2002).
7. Saderi, N. et al. The NPY intergeniculate leaflet projections to the suprachiasmatic nucleus transmit metabolic conditions. *Neuroscience* **246**, 291–300 (2013).
8. Moore, R. Y. & Card, J. P. Intergeniculate leaflet: an anatomically and functionally distinct subdivision of the lateral geniculate complex. *J. Comp. Neurol.* **344**, 403–430 (1994).
9. Morin, L. P. Neuroanatomy of the extended circadian rhythm system. *Exp. Neurol.* **243**, 4–20 (2013).
10. Wams, E. J., Riede, S. J., van der Laan, I., ten Bulte, T. & Hut, R. A. in *Biological Timekeeping: Clocks, Rhythms and Behaviour* (ed. Kumar, V.) 395–404 (Springer India, 2017).
11. Chew, K. S. et al. A subset of ipRGCs regulates both maturation of the circadian clock and segregation of retinogeniculate projections in mice. *eLife* **6**, e22861 (2017).
12. Mistlberger, R. E. Food-anticipatory circadian rhythms: concepts and methods. *Eur. J. Neurosci.* **30**, 1718–1729 (2009).
13. Acosta-Galvan, G. et al. Interaction between hypothalamic dorsomedial nucleus and the suprachiasmatic nucleus determines intensity of food anticipatory behavior. *Proc. Natl Acad. Sci. USA* **108**, 5813–5818 (2011).
14. Patton, D. F. et al. Photoc and pineal modulation of food anticipatory circadian activity rhythms in rodents. *PLoS ONE* **8**, e81588 (2013).
15. Nakazato, M. et al. A role for ghrelin in the central regulation of feeding. *Nature* **409**, 194–198 (2001).
16. Camiña, J. P. et al. Regulation of ghrelin secretion and action. *Endocrine* **22**, 5–12 (2003).
17. Blum, I. D., Lamont, E. W., Rodrigues, T. & Abizaid, A. Isolating neural correlates of the pacemaker for food anticipation. *PLoS ONE* **7**, e36117 (2012).
18. Glass, J. D., Guinn, J., Kaur, G. & Francis, J. M. On the intrinsic regulation of neuropeptide Y release in the mammalian suprachiasmatic nucleus circadian clock. *Eur. J. Neurosci.* **31**, 1117–1126 (2010).
19. Andermann, M. L. & Lowell, B. B. Toward a wiring diagram understanding of appetite control. *Neuron* **95**, 757–778 (2017).
20. Chen, S.-K., Badea, T. C. & Hattar, S. Photoentrainment and pupillary light reflex are mediated by distinct populations of ipRGCs. *Nature* **476**, 92–95 (2011).
21. Fernandez, D. C. et al. Light affects mood and learning through distinct retina–brain pathways. *Cell* **175**, 71–84.e18 (2018).
22. Sindelar, D. K., Palmiter, R. D., Woods, S. C. & Schwartz, M. W. Attenuated feeding responses to circadian and palatability cues in mice lacking neuropeptide Y. *Peptides* **26**, 2597–2602 (2005).
23. May, P. J. The mammalian superior colliculus: laminar structure and connections. *Prog. Brain Res.* **151**, 321–378 (2006).
24. Gandhi, N. J. & Katnani, H. A. Motor functions of the superior colliculus. *Annu. Rev. Neurosci.* **34**, 205–231 (2011).
25. Stein, B. E., Stanford, T. R. & Rowland, B. A. Development of multisensory integration from the perspective of the individual neuron. *Nat. Rev. Neurosci.* **15**, 520–535 (2014).
26. Mundiñano, I. C. & Martínez-Millán, L. Somatosensory cross-modal plasticity in the superior colliculus of visually deafferented rats. *Neuroscience* **165**, 1457–1470 (2010).
27. Waterson, M. J. & Horvath, T. L. Neuronal regulation of energy homeostasis: beyond the hypothalamus and feeding. *Cell Metab.* **22**, 962–970 (2015).
28. Sheng, Z., Santiago, A. M., Thomas, M. P. & Routh, V. H. Metabolic regulation of lateral hypothalamic glucose-inhibited orexin neurons may influence midbrain reward neurocircuitry. *Mol. Cell. Neurosci.* **62**, 30–41 (2014).
29. Azevedo, E. P. et al. A role of Drd2 hippocampal neurons in context-dependent food intake. *Neuron* **102**, 873–886.e5 (2019).
30. Challet, E. The circadian regulation of food intake. *Nat. Rev. Endocrinol.* **15**, 393–405 (2019).
31. Pendergast, J. S. & Yamazaki, S. The mysterious food-entrainable oscillator: insights from mutant and engineered mouse models. *J. Biol. Rhythms* **33**, 458–474 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

When appropriate, statistical methods were used to predetermine sample size. The experiments were randomized and investigators were blinded to allocation during experiments and outcome assessment.

Mice

Female and male mice were used in this study. Wild-type mice of a mixed background (B6/129 F₁ hybrid, stock no. 101043), *Npy*^{tm1Rpa} (stock no. 4545) and *Npy*^{cre} (stock no. 27851) mice were obtained from the Jackson Laboratory. *Opn4*^{DTA/DTA}, *Opn4*^{attnDTA/attnDTA} and *Opn4*^{cre/+} *Brn3b*^{DTA/+} mouse lines have previously been described^{4,11,20,21}. All mice were handled in accordance with guidelines of the Animal Care and Use Committees of the National Institute of Mental Health (NIMH). All efforts were made to minimize the pain and the number of mice used.

Locomotor and feeding activity measurements

Mice were housed under a 12 h:2 h light:dark cycle or constant darkness at a temperature of 22 °C. During all the behavioural experiments, mice were single-housed. General locomotor activity was monitored using infrared motion detectors from Mini Mitter (Respironics) mounted on top of the cages. Data was collected in 5-min bins using Vital-View software (Mini Mitter). Feeding activity was monitored using programmable feeders (Actimetrics), as previously described³². ClockLab (Actimetrics) software was used to set the TRF schedules, and to measure number of pellets consumed. Dustless Precision Pellets were used (300 mg pellets; Bio-Serv; product no. F0170). Actograms, total activity, periodograms and period lengths were obtained and calculated using ClockLab (Actimetrics).

The locomotor activity (9 h) before food access was measured and results were expressed as percentage of activity relative to total locomotor activity; the area under the curve was analysed for all mice.

Food-anticipatory activity was determined as locomotor activity measured 3 h before food access relative to the total activity during the TRF protocol.

When exposed to constant darkness with ad libitum access to food, we observed that some genetic mouse lines displayed different free-running periods. Therefore, we implemented a graded-score analysis to account for potential variations in the food-anticipatory activity caused by differences in the free-running periods displayed before and during the TRF paradigm. The analysis was performed with the experimenter blind to genotype and/or condition. Entrainment to TRF was graded on a scale from 0 to 5: a score of 0 indicates no food-anticipatory activity (defined as less than 5% of total activity), with unperturbed free-running locomotor activity (defined as less than 15-min change in period length). A score of 0.1–1.0 indicates no food-anticipatory activity (defined as less than 5% of total activity), with changes in the free-running locomotor activity (defined as more than 15-min change in period length). A score of 1.1–2.0 indicates a weak or sporadic food-anticipatory activity (defined as less than 10% of total activity), and with or without changes in free-running locomotor activity. A score of 2.1–3.0 indicates a sustained food-anticipatory activity (defined as 10–15% of total activity), with or without changes in free-running locomotor activity. A score of 3.1–4.0 indicates robust food-anticipatory activity (defined as more than 15% of total activity) and unperturbed free-running locomotor activity (defined as less than 15-min change in period length). A score of 4.1–5.0 indicates robust food-anticipatory activity (defined as more than 15% of total activity), with changes in the free-running locomotor activity (defined as more than 15-min change in period length).

Retinal injections

Retinal projections were visualized using intravitreal injections (1 µl) of the tracer cholera toxin b-subunit (CTB) fluorescently conjugated (to Alexa Fluor 488 or 594, Thermofisher). Mice were anaesthetized using isoflurane and placed under a stereo-microscope. The microscope and all the instruments were properly cleaned and sterilized. A glass needle (pulled

10-µl microcapillary tube, Sigma P0674) and a 10-µl Hamilton syringe were used to drive the solution into the vitreous chamber of the eye to ensure delivery specifically to the retina. After slowly injecting the total volume, pipette was left in place for 60–90 s. Mice recovered from injections on a heating pad until they woke from anaesthesia. After injections, mice were given a 3–4-day recovery period. Finally, mice were deeply anaesthetized, and perfused intracardially with 4% paraformaldehyde (Electron Microscopy Sciences). Brains were post-fixed overnight in the same fixative, and coronal brain sections were obtained using a cryostat.

Stereotaxic injections and optical fibre implantation

The stereotaxic frame and all instruments were properly cleaned and sterilized. Mice were deeply anaesthetized using isoflurane, as confirmed by complete absence of flinching response to pinch. Skull fur was shaved, the head of the mouse was then fixed to the stereotaxic frame, cleaned by scrubbing with povidone-iodine and 70% ethanol and the skull was exposed using a sterile scalpel. A small hole was drilled over the region of interest. Coordinates follow the Paxinos and Franklin mouse atlas³³. For IGL injections, the following coordinates were used: –4.21 mm from bregma, ±2.45 mm lateral from midline and –2.30 mm vertical from cortical surface. AAV injections were performed using a microinjector (Nanojector II, Drummond Scientific) and pulled 10-µl microcapillary pipettes. During the entire procedure, a heating pad was used to maintain stable body temperature in mice. At the end of the surgical procedure, the incision was closed using nylon sutures. Systemic analgesics (either buprenorphine, 0.1 mg/kg, or meloxicam, 1 mg/kg) were administered before and after surgery.

For anatomical analysis, AAVs (AAV2/9-phSyn1(S)-Flex-tdTomato-T2A-SynEGFP-WPRE obtained from Boston Children's Hospital Viral Core with a titre of 4.26×10^{13} genome copies (GC)/ml and AAV5/Syn-DIO-hChR2(H134R)-EGFP-WPRE-HGHpA obtained from Addgene no. 20298 with a titre of 1×10^{13} GC/ml) were used. Mice were perfused at different times after injection, and the brains were subsequently sectioned on a cryostat.

For silencing IGL^{NPY} neurons, AAVs (AAV5/Syn-DIO-hChR2(H134R)-EGFP-WPRE-HGHpA obtained from Addgene no. 20298 with a titre of 1×10^{13} GC/ml (control), and pAAV5/CMV-DIO-eGFP-2A-TeNT, GVVC-AAV-71 (TeNT) obtained from Stanford University no. 2237 with a titre of 1×10^{13} GC/ml) were used. Mice were then tested for TRF.

For optogenetic experiments, AAVs (AAV5/DIO-tdTomato (control for the optogenetic virus) and AAV5/DIO-ArchT-tdTomato obtained from Addgene) were bilaterally injected into the IGL. A week after virus injections, optical fibres (100-µm diameter, Thorlabs) were implanted above the SCN (–0.50 mm from bregma, ±0.15 mm lateral from midline and –5.60 mm vertical from cortical surface, 9.99° angle) and were affixed to the skull using Metabond Cement System (Parkell) and Jet Brand dental acrylic (Lang Dental Manufacturing). Following all surgical procedures, mice recovered on a heating pad and returned to their home cages after 24-h post-surgery recovery and monitoring. Mice received subcutaneous injections of meloxicam (1–2 mg/kg) for analgesia and anti-inflammatory purposes. Two weeks after recovery under light:dark cycle, mice were exposed to TRF for 3–4 weeks.

The optical fibres were connected to a laser source (Ce:YAG, Ce:YAG & LED Driver, Doric Lenses) via a dual fibre rotary joint (FRJ_1x2i_FC-2FC; Doric Lenses) using an optic fibre sleeve (Thorlabs). The light intensity at the interface between the fibre tip and the mouse was 10 mW. Optical stimulation was delivered 2 h before food delivery, by applying 3 pulses of 20 min of light, with 20-min intervals. Mice without correct targeting of tracers and/or vectors were excluded from this study.

Eye enucleation

P0 mice were deeply anaesthetized, and an approximately 1-mm incision was made across each eyelid using a sterile scalpel. Finally, sterilized forceps were used to pull the eyes free of the orbitals. For adult mice enucleation, mice were deeply anaesthetized and a sterile curved scissor was used to cut the optic nerve and remove both eyes. Bleeding

Article

was controlled by orbital pressure on the eye with a sterile cotton swab. Before and after surgery, systemic analgesics (buprenorphine, 0.1 mg/kg) were administered. Mice were monitored over the next several days for signs of infection.

Immunofluorescence

Brain sections were incubated in 0.1 M PBS with 3% goat serum (Vector Labs) and 0.3% Triton X-100 (Sigma Aldrich) for 2 h, and then incubated using the following antibodies (overnight, at 4 °C): rabbit anti-RFP (MBL PM005, 1:1,000); chicken anti-GFP (AbCam Ab13970, 1:2,000); mouse IgG1 anti-FOS (EnCor MCA-2H2, 1:1,000); rabbit antibody anti-NPY (Peninsula lab T-4070, 1:500). After several washing steps, Alexa-Fluor-conjugated secondary antibodies were used (Molecular Probes, 1:500, for 2 h at room temperature). Finally, slides were mounted using AntiFade medium (Molecular Probes). Images were acquired using an Eclipse Ti2 confocal microscope (Nikon).

Quantification and statistical analysis

For all morphometric image processing, digitalized captured TIF images were assembled and processed with NIS Elements (Nikon) Version 5.02 and Adobe Photoshop (Adobe Systems), and transferred to ImageJ software (NIH). Sample analysis was performed with the experimenter blind to condition. All the nomenclature used in the Article follows that of Paxinos and Franklin Atlas³³.

Morphometric analysis of IGL and arcuate nucleus

Total IGL or arcuate nucleus areas of analysis were manually outlined in coronal brain sections on the basis of DAPI staining. Bilateral nuclei were evaluated per section.

FOS induction. Mice were exposed to the TRF paradigm for three weeks. On the 21st day, mice were perfused immediately before the time of food delivery. As control group, mice with ad libitum access to food were perfused at circadian time 11–12 (before activity onset). In all cases, mice were under constant darkness conditions. FOS⁺ cells were manually counted in the delineated IGL or arcuate nucleus area, and results obtained from 5–6 separate coronal sections were averaged per mouse.

NPY levels. Mice housed under a 12 h:2 h light:dark cycle were perfused at circadian time 14–16. Digital images were converted to 8-bit greyscale, and the optic density indicating NPY expression levels in the IGL was measured. For measuring number of NPY⁺ neurons in the IGL, a 3D reconstruction was obtained from z-stack images (15–20 µm), which were overexposed (to account for different NPY expression levels) and the number of NPY⁺ somas was manually counted. Finally, the number of DAPI⁺ nuclei was manually counted. In all cases, results obtained from 5–6 separate coronal sections were averaged per mouse.

SCN morphometric analysis

Total SCN area of analysis was manually outlined in coronal brain sections on the basis of DAPI staining. Digital images were converted to 8-bit greyscale, and the optic density indicating NPY expression or Syn-GFP levels was measured. Results obtained from 3–6 separate coronal sections were averaged per mouse.

Metabolic measurements

Anorexigenic and orexigenic hormones were measured by enzyme-linked immunosorbent assay (ELISA). Blood samples were

collected and stored in EDTA-coated tubes (BD Microtainer 365974). Plasma was then obtained, and total ghrelin, insulin and leptin levels were measured using the following ELISA kits: rat/mouse total ghrelin (Millipore EZRGRT-91K); rat insulin (Crystal Chem 90010); mouse leptin (R&D Systems MOB00). Blood glucose was measured using a regular blood glucometer. Samples were taken from the mice tails, under dim red light.

Body composition was measured in awake mice using quantitative magnetic resonance technology (EchoMRI composition analyser).

Statistical analysis

Calculation of sample size per experiment was determined, or confirmed by post hoc analyses, using G*Power 3 software^{34,35}.

Statistical analysis of results was performed using Student's *t*-test (parametric or non-parametric (Mann–Whitney)), or analysis of variance (ANOVA), followed by Tukey's or Sidak's multiple comparisons tests, as stated. All the analyses were done using GraphPad Prism, version 7.0a.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The principal data supporting the findings of this Article are available within the figures and the Supplementary Information; additional data that support the findings of this study are available from the corresponding authors on request.

32. Acosta-Rodríguez, V. A., de Groot, M. H. M., Rijo-Ferreira, F., Green, C. B. & Takahashi, J. S. Mice under caloric restriction self-impose a temporal restriction of food intake as revealed by an automated feeder system. *Cell Metab.* **26**, 267–277.e2 (2017).
33. Franklin, K. B. J. & Paxinos, G. *Paxinos and Franklin's The Mouse Brain in Stereotaxic Coordinates* (Academic, 2013).
34. Charan, J. & Kantharia, N. D. How to calculate sample size in animal studies? *J. Pharmacol. Pharmacother.* **4**, 303–306 (2013).
35. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).

Acknowledgements We thank the members of the SLCR at NIMH, the Johns Hopkins Biology Mouse Tri-Lab and M. E. Mercu for helpful discussions; O. Gavrilova and the NIDDK Mouse Metabolism Core for their skillful technical assistance; and V. Acosta-Rodríguez for her assistance with the programmable feeders. This work was supported by the NIH (GM076430, EY027202), the generous contributions of the PEW Charitable Trusts (to D.C.F.) and the intramural research fund at the National Institute of Mental Health (ZIA MH002964-02).

Author contributions D.C.F. contributed to conceptualization, formal analysis, investigation, methodology, project administration, supervision, visualization and writing (original draft and editing); R.K., J.L. and J.M. contributed to investigation, methodology and writing (reviewing and editing). P.Q.D. contributed to investigation and methodology. M.P. and H.Z. contributed to funding acquisition and writing (review and editing). S.H. contributed to conceptualization, funding acquisition, project administration, supervision and writing (original draft and editing).

Competing interests The authors declare no competing interests.

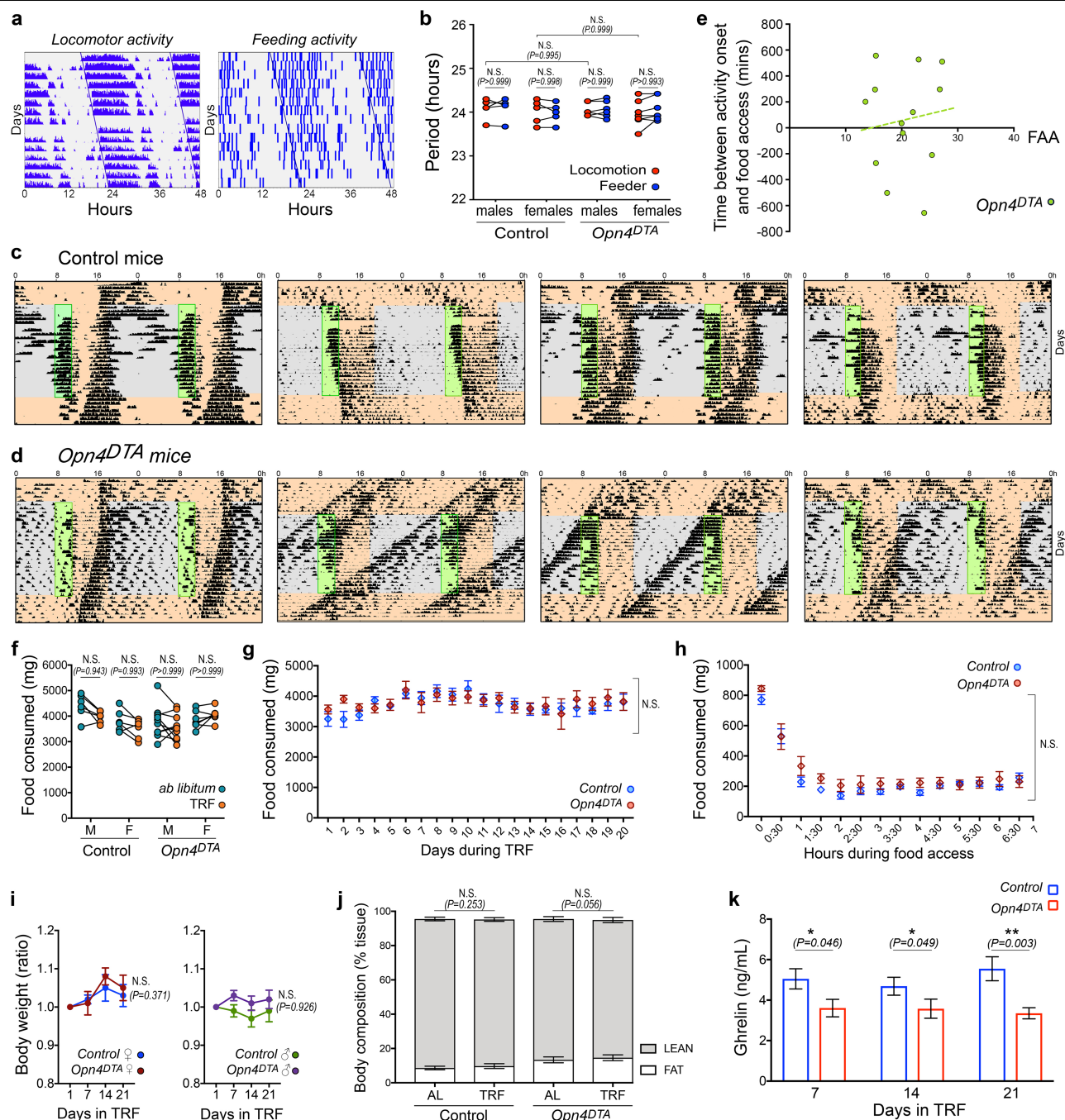
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2204-1>.

Correspondence and requests for materials should be addressed to D.C.F. or S.H.

Peer review information Nature thanks Joseph Bass, Sarah Chellappa, Frank Scheer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

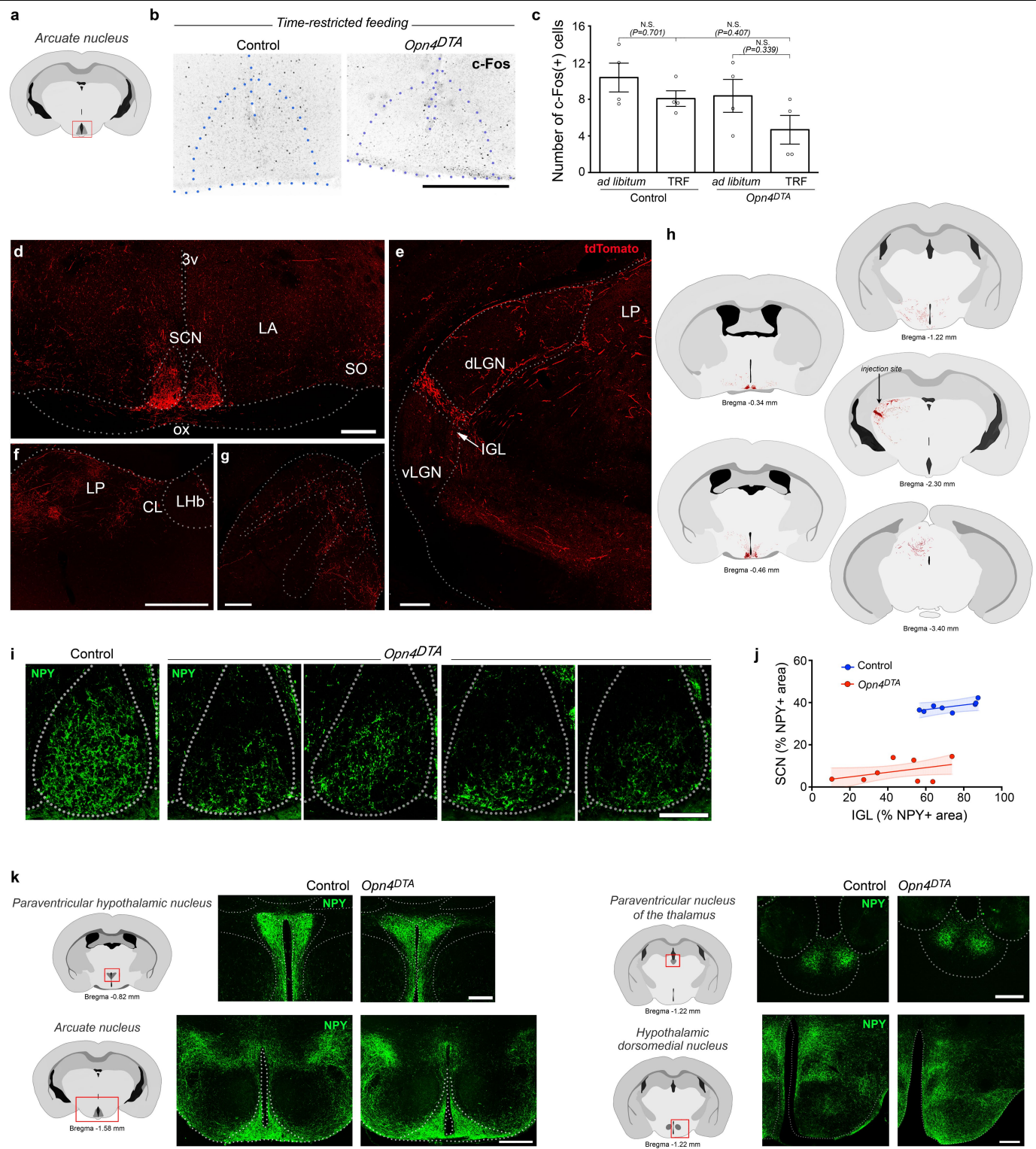
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Early ablation of ipRGCs alters entrainment to TRF.

a, b, General locomotor activity and feeding behaviour were monitored using infrared sensors and programmable feeders, respectively. Representative actograms obtained from an *Opn4^{DTA}* mouse under free-running (constant darkness and ad libitum access to food) are shown (**a**). Periodograms were obtained, and no differences in period lengths were found between locomotor activity and the feeding behaviour for both groups (**b**). Data are mean \pm s.e.m. ($n = 12$ mice for each genotype), two-way ANOVA, followed by Sidak's multiple comparisons test. **c, d**, Representative actograms obtained from control (**c**) and *Opn4^{DTA}* (**d**) mice exposed to TRF are shown. **e**, Under constant darkness and ad libitum access to food, *Opn4^{DTA}* mice displayed different free-running periods. Therefore, we analysed whether there is any correlation between the food-anticipatory activity (measured during a 3-h time window) and the time difference (measured in minutes) between the onset of locomotor activity and the time of food delivery measured during the first day of food restriction. No significant correlations (Pearson correlation test, $P = 0.6379$) were observed in *Opn4^{DTA}* mice ($n = 13$ mice). **f**, The daily total amount of food consumed was

measured in male (M) and female (F) control and *Opn4^{DTA}* mice exposed to the free-running (ad libitum access to food) or TRF paradigm. Data are mean \pm s.e.m. ($n = 8$ mice for each genotype), two-way ANOVA, followed by Sidak's multiple comparisons test. **g**, Total food consumption per day during TRF. Data are mean \pm s.e.m. ($n = 8$ mice for each genotype), multiple Student's *t*-test, two tailed. **h**, Pattern of food consumption during the 7 h of food access. Data are mean \pm s.e.m. ($n = 8$ mice for each genotype); multiple Student's *t*-test, two tailed. **i**, Measurement of the body weight of female and male mice exposed to the TRF paradigm. Data are mean \pm s.e.m. ($n = 12$ mice for each genotype), two-way ANOVA, followed by Sidak's multiple comparisons test. **j**, The body composition was measured in mice with ad libitum access to food (AL), or on the 21st day of TRF. Data are mean \pm SEM ($n = 8$ mice for each genotype), two-way ANOVA, followed by Sidak's multiple comparisons test. **k**, Total ghrelin levels (ng mL⁻¹) were measured in control and *Opn4^{DTA}* mice after 7, 14, and 21 days of TRF. In all cases, samples were collected immediately before food delivery. Data are mean \pm SEM ($n = 8$ mice for each genotype), Student's *t*-test, two tailed.

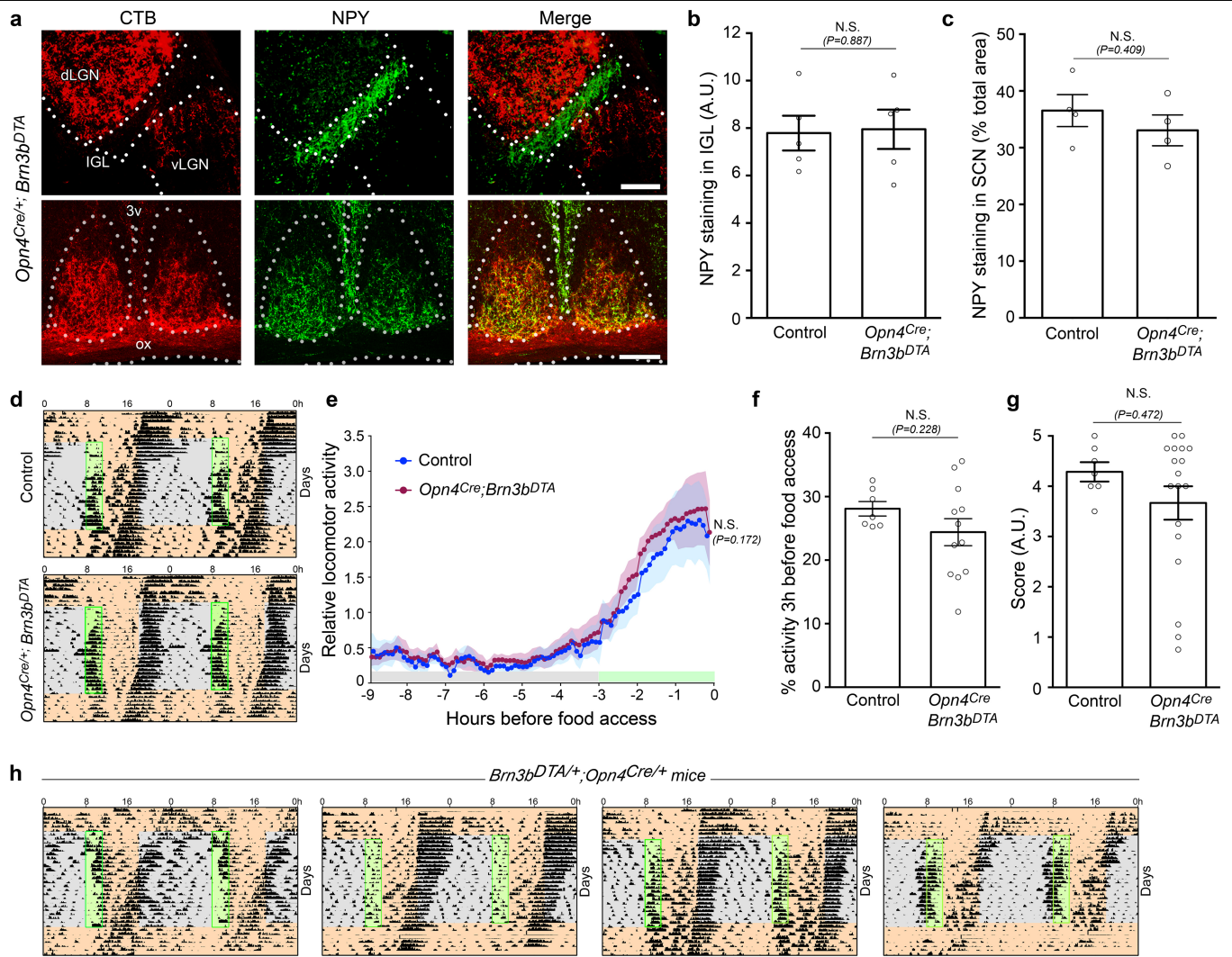


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Early ablation of ipRGCs causes alterations in the IGL^{NPY}-SCN circuit. **a–c**, FOS induction in the arcuate nucleus mediated by the expected food access. Mice were exposed to TRF and perfused on the 21st day at the expected food time (immediately before food delivery). As controls, mice with *ad libitum* access to food were perfused at circadian time 12. All mice were housed under constant darkness. The area analysed is shown in a diagram of a representative coronal brain section (**a**). Representative images of control and *Opn4^{DTA}* mice exposed to TRF are shown (**b**); the number of FOS⁺ cells in the arcuate nucleus was quantified (**c**). Data are mean \pm s.e.m. ($n = 4$ mice for each genotype), two-tailed Tukey's test. **d–h**, Projection pattern of IGL^{NPY} cells. *Npy^{cre/+}* mice were unilaterally injected in the IGL using a Cre-dependent AAV encoding tdTomato (AAV2/9-phSyn1(S)-Flex-tdTomato-T2A-SynEGFP-WPRE). IGL^{NPY} neurons send dense and bilateral projections to the SCN (**d**) and, to a least extent, unilateral projections to other brain targets, including the dorsal geniculate and dorsal thalamus (**e** and **f**, respectively), and the superior colliculus (**g**). The complete pattern of IGL^{NPY} projections is shown in a diagram of representative coronal brain sections (**h**). Three independent experiments were performed with similar results. **i**, Representative SCN sections obtained

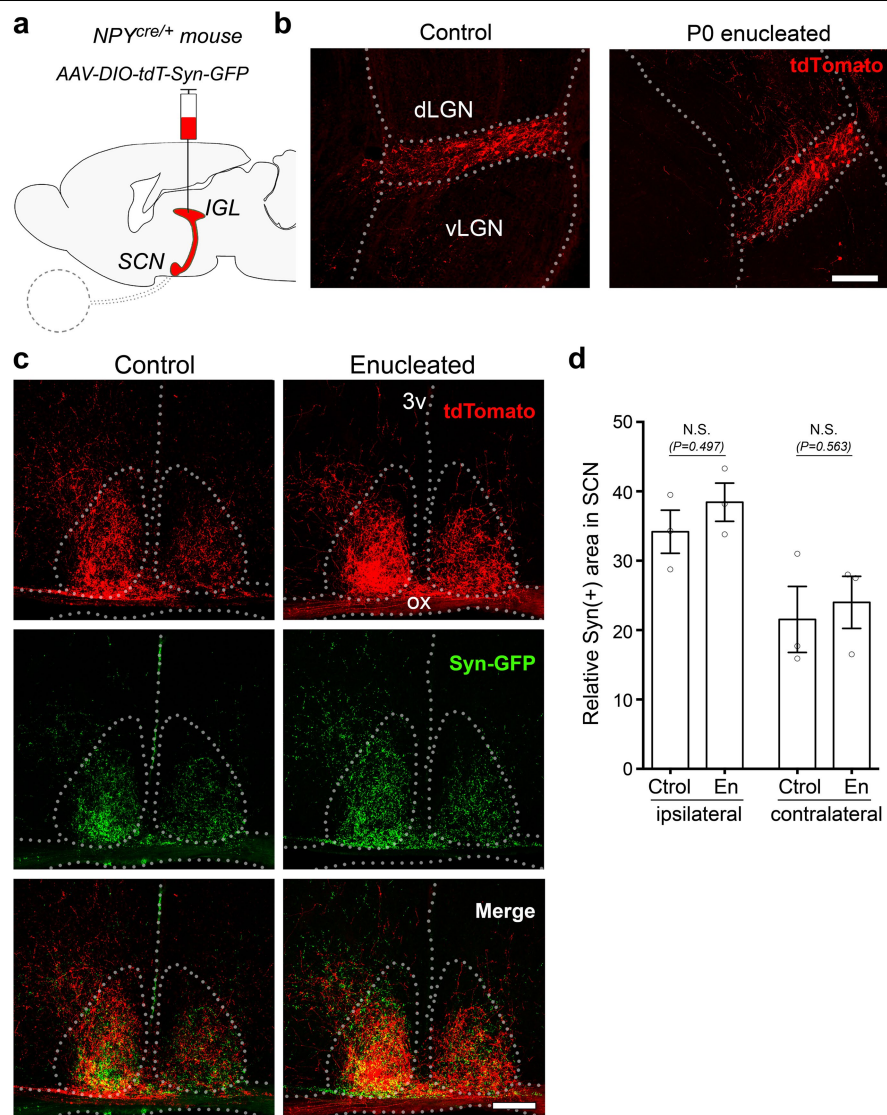
from control and *Opn4^{DTA}* mice are shown. Marked alterations in the pattern of NPY staining in the SCN were observed in *Opn4^{DTA}* mice, compared to control mice. **j**, Correlation between NPY level in somas and axonal terminals, measured in the IGL and SCN, respectively. Results obtained from control and *Opn4^{DTA}* mice are shown. Pearson r values were measured for both groups (control = 0.728; *Opn4^{DTA}* = 0.389). A linear regression was applied, and the comparison of slope fits was not significantly different (slope \pm s.e., control = 0.136 ± 0.052 ; *Opn4^{DTA}* = 0.100 ± 0.096). The asymptotic normal 95% confidence interval is shown for both groups. ($n = 7$ control mice, 8 *Opn4^{DTA}* mice). **k**, Brain targets that are not innervated by ipRGCs and that express NPY were studied in control and *Opn4^{DTA}* mice. No obvious changes in NPY expression levels were observed in the paraventricular hypothalamic nucleus, arcuate nucleus, paraventricular nucleus of the thalamus or hypothalamic dorsomedial nucleus. Three independent experiments were performed with similar results. CL, centrolateral nucleus of the thalamus; LA, lateroanterior hypothalamic nucleus; LHb, lateral habenula LP, lateral posterior thalamic nucleus; SO, supraoptic nucleus. Scale bar, 100 μ m (**b**, **i**, **k**), 200 μ m (**d**, **f**), 400 μ m (**e**, **g**).

shown. NPY* fibres were measured in the SCN area. Five independent experiments were performed with similar results. **d–g**, Representative actograms obtained from wild-type mice enucleated at P0 (**d**), P40 (**e**) or P90 (**f**) under TRF. A score analysis was performed for all actograms obtained (**g**). Data are mean \pm SEM ($n = 5$ control, 6 En P0, 11 En P40 and 8 En P90 mice), Student's *t* non-parametric (Mann–Whitney) test, two-tailed. Scale bar, 100 μ m.



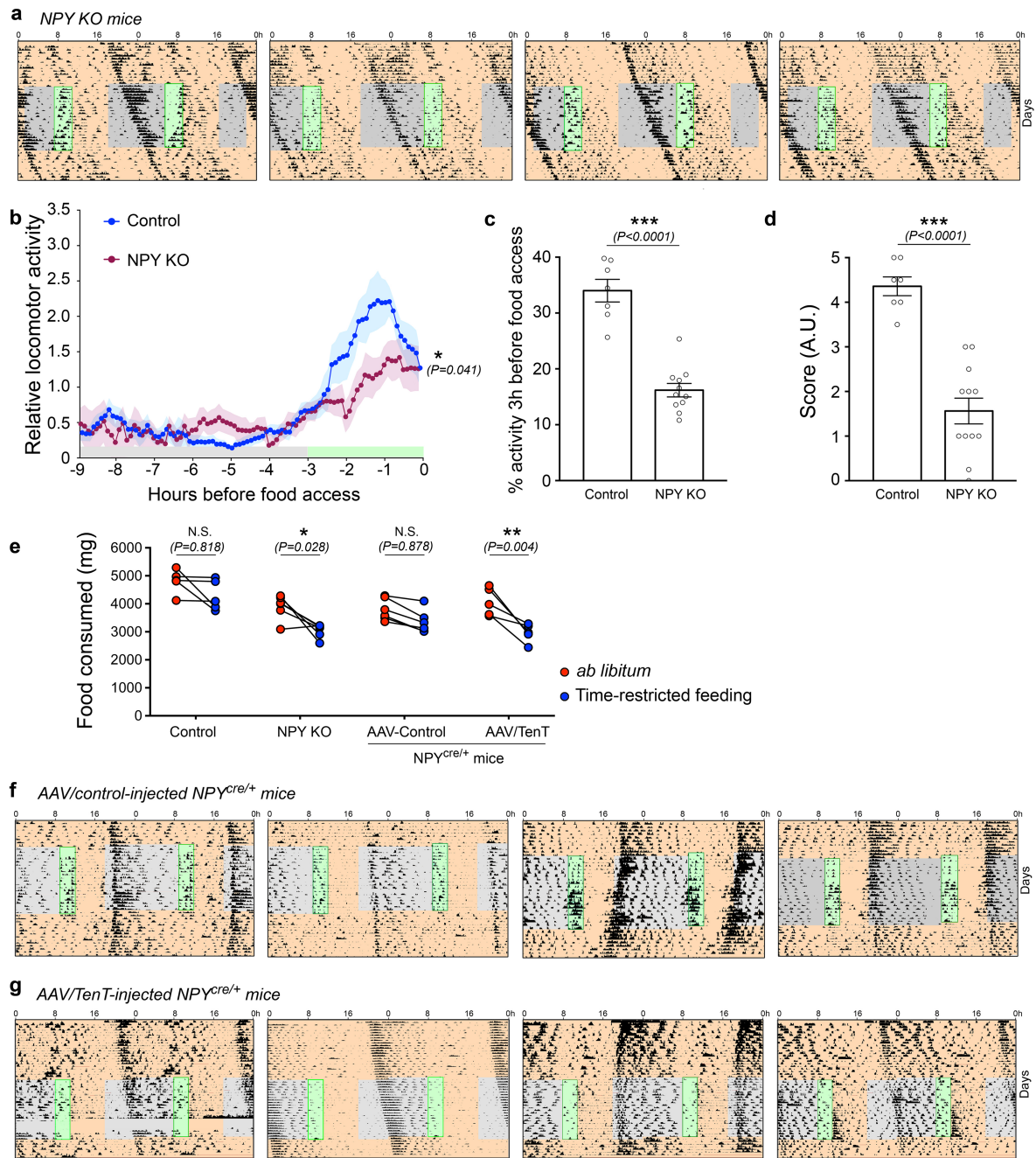
Extended Data Fig. 4 | Role of SCN-projecting ipRGCs in nonphotic entrainment. **a–c**, Retinal innervation (CTB) (red) and NPY staining (green) in the IGL and SCN in control and *Opn4^{Cre/+} Brn3b^{DTA}/+* mice. Representative coronal sections are shown (**a**). NPY staining in the IGL (**b**) and SCN (**c**) were analysed in three-month-old mice. Data are mean \pm SEM ($n=5$ mice for each genotype), two-tailed Student's *t*-test. **d–h**, Control and *Opn4^{Cre/+} Brn3b^{DTA}/+* mice were exposed to TRF. Representative actograms are shown (**d**). The locomotor

activity before food access (**e**) and the food-anticipatory activity (**f**) were measured. Data are mean \pm s.e.m. ($n=7$ control mice, 12 *Opn4^{Cre/+} Brn3b^{DTA}/+* mice), by two-tailed Student's *t*-test. A score analysis was performed for all actograms obtained (**g**). Data are mean \pm s.e.m. ($n=7$ control mice, 18 *Opn4^{Cre/+} Brn3b^{DTA}/+* mice), by Student's *t* non-parametric (Mann–Whitney) test, two-tailed. Representative actograms obtained from *Opn4^{Cre/+} Brn3b^{DTA}/+* mice under TRF (**h**). Scale bars, 100 μ m (**a**, bottom), 200 μ m (**a**, top).



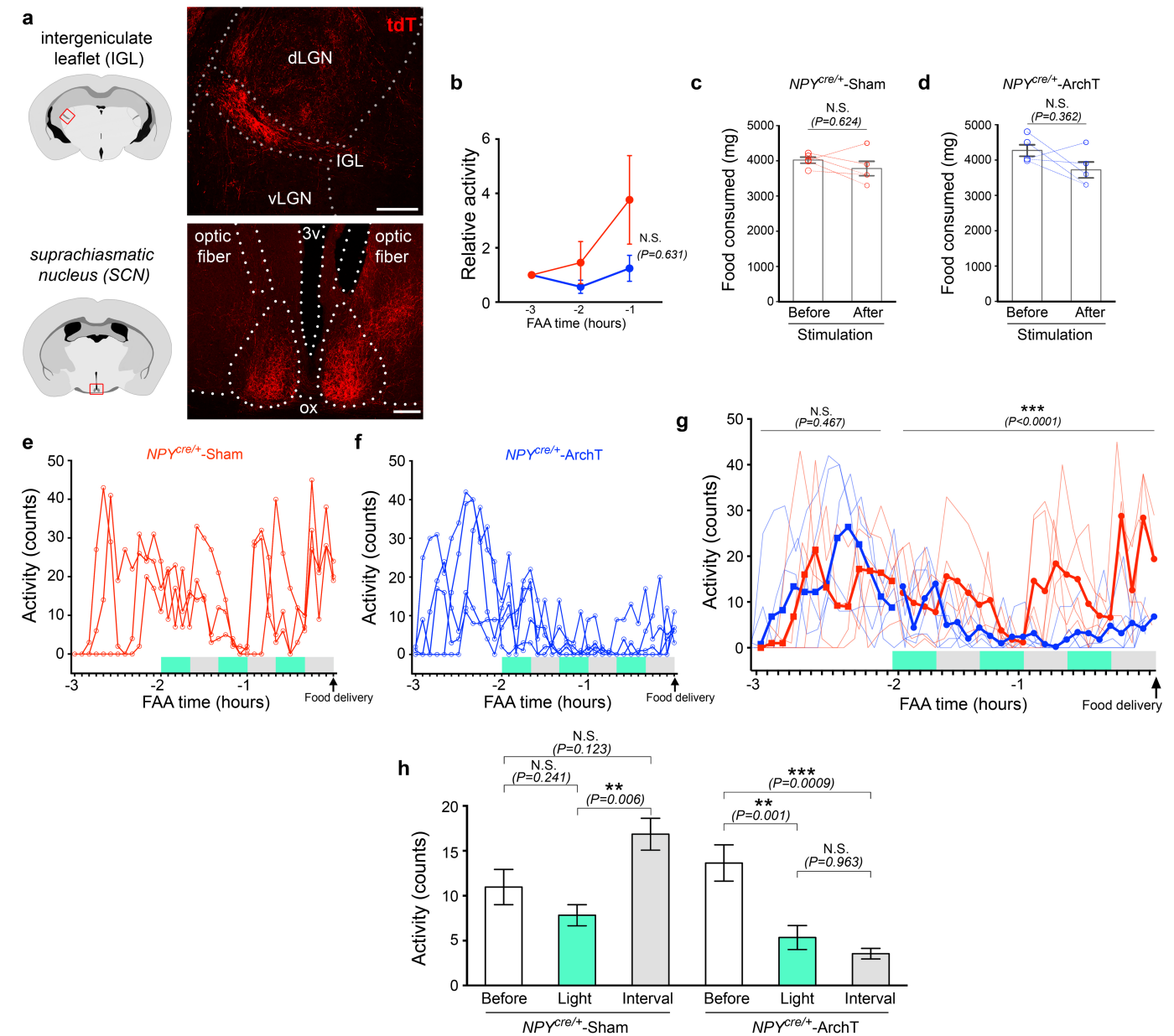
Extended Data Fig. 5 | IprGC input to SCN during early postnatal stages influences afferent IGL^{NPY} projections. **a**, *Npy^{cre/+}* mice, with or without bilateral enucleation at P0, were injected in the IGL with Cre-dependent AAV-tdT-Syn-GFP. **b**, Representative IGL injections in adult control and P0 enucleated *Npy^{cre/+}* mice are shown. Three independent experiments were

performed with similar results. **c**, **d**, Axonal projections (tdTomato) (red) from IGL^{NPY} cells and their synaptic terminals (syn-GFP) (green) are shown (**c**). IGL^{NPY} synaptic terminals were quantified (**d**). Data are mean ± s.e.m. ($n = 3$ mice for each condition), two-tailed Student's *t*-test. Scale bars, 200 μ m (**b**), 100 μ m (**c**).



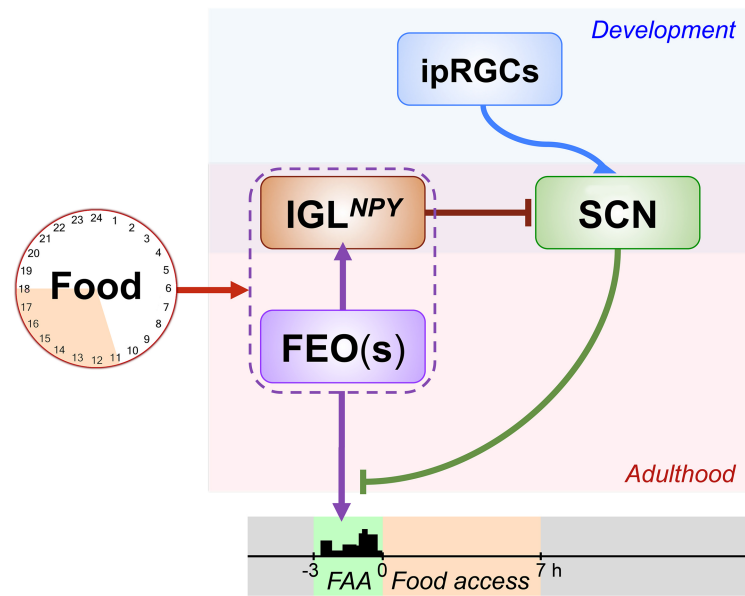
Extended Data Fig. 6 | NPY signalling in the IGL-SCN circuit controls nonphotic entrainment. **a–d**, Representative actograms obtained from NPY-knockout (KO) (*Npy^{cre/cre}*) mice exposed to TRF are shown (**a**). The locomotor activity 9 h before food access (**b**) and the food-anticipatory activity (**c**) were measured for control and NPY-knockout mice. Data are mean \pm s.e.m. ($n=7$ control mice, 11 NPY-knockout mice), two-tailed Student's *t*-test. Additionally, a score analysis was performed for all actograms obtained (**d**). Data are mean \pm s.e.m. ($n=7$ control mice, 11 NPY-knockout mice), Student's *t* non-parametric (Mann-Whitney) test, two-tailed. In addition, a second mouse line (*Npy^{tm1Rpa}*) was used to evaluate the effects of NPY ablation. Results obtained from both NPY-knockout mouse lines were indistinguishable (data

not shown). **e**, The daily total amount of food consumed (during ad libitum access to food and TRF) was measured in control, NPY-knockout (*Npy^{cre/cre}*), and *Npy^{cre/+}* mice bilaterally injected in the IGL with a control AAV (AAV5/Syn-DIO-hChR2(H134R)-EGFP-WPRE-HGHpA) or AAV encoding TenT (pAAV5/CMV-DIO-eGFP-2A-TenT). Under free-running conditions (constant darkness and ad libitum access to food), the amount of food consumed by control and NPY-knockout mice ($P=0.0267$), as well as control and *Npy^{cre/+}* mice ($P=0.008$), was significantly different. Data are mean \pm s.e.m. ($n=5$ mice for each condition), two-way ANOVA, followed by Sidak's multiple comparisons test. **f**, **g**, Representative actograms obtained from *Npy^{cre/+}* mice exposed to TRF and injected with control AAV (**f**) or AAV encoding TenT (**g**) are shown.



Extended Data Fig. 7 | Neural silencing of the IGL^{NPY}-SCN circuit. **a**, Sites of injections (IGL) and optical fibre implantation (SCN) were confirmed at the end of the experiments. **b** *Npy^{cre/+}* sham (AAV/DIO-tdTomato) and *Npy^{cre/+}* ArchT (AAV5/DIO-ArchT-tdTomato) mice were housed under TRF for 3 weeks and the locomotor activity was then measured for 3 h before food access. Results are shown as relative activity (total counts per hour) during food-anticipatory activity. Data are mean \pm s.e.m. ($n = 5$ mice for each condition), two-way ANOVA, followed by Sidak's multiple comparisons test. **c, d**, Food consumed during TRF. The average amount of food consumed was measured for 10 days before optogenetic stimulation (before); the amount of food consumed the day of neural silencing was also measured (after) for both groups of mice. Data are mean \pm s.e.m. ($n = 5$ mice for each condition), paired Student's *t*-test, two-tailed. **e–h**, Locomotor activity was quantified (in 5-min bins) starting 3 h before food access. Two hours before food access, neural silencing was induced by applying 3 pulses of light, of 20 min each (shown in green), with 20-min intervals (shown in grey). Results of locomotor activity obtained for all *Npy^{cre/+}* sham (**e**) and

Npy^{cre/+} ArchT (**f**) mice are shown. The average of results is also shown (**g**); similar food-anticipatory activity was observed in both groups without optical stimulation, whereas a significant reduction in locomotor activity was found in *Npy^{cre/+}* ArchT mice during optogenetic stimulation. Data are mean \pm s.e.m. ($n = 5$ mice for each condition), two-way ANOVA, followed by Sidak's multiple comparisons test. For *Npy^{cre/+}* sham mice, no changes in total activity were observed during light stimulation (light), or intervals between light pulses (interval), compared with activity measured during the first hour of food-anticipatory activity (before) (**h**). We observed a significant reduction in activity with light stimulation compared with interval periods, suggesting that the light could have a direct masking effect on mouse activity. However, the activity was fully recovered immediately after stimulation. *Npy^{cre/+}* ArchT mice displayed similar and reduced activity for both stimulation and interval periods, compared with the activity displayed before optogenetic stimulation. Data are mean \pm s.e.m. ($n = 5$ mice for each condition), two-tailed Tukey's test. Scale bars, 100 μ m (**a**, bottom), 200 μ m (**a**, top).



Extended Data Fig. 8 | Putative model for the circuit that drives circadian food-anticipatory activity. Time-restricted access to food constitutes a strong environmental cue that causes the alignment of the circadian system to feeding schedules, driving food-anticipatory activity that precedes the expected meal. The current view in the field is that a widespread system—composed of elements referred to as food-entrainable oscillators (FEOs)—controls the physiological and behavioural responses to TRF. Among the candidates for FEOs are areas of the hypothalamus (such as the paraventricular nucleus, ventromedial hypothalamic nucleus, dorsomedial hypothalamic nucleus and arcuate nucleus), thalamic areas (such as the paraventricular thalamus and the IGL), the brainstem (including the dorsal raphe nucleus and parabrachial nucleus), other brain regions (such as the dorsal striatum, infralimbic cortex, nucleus accumbens and cerebellum) as well as peripheral targets (such as the gastrointestinal system). In this Article, we have delineated

a brain circuit (IGL^{NPY}–SCN) that is critical for driving food-anticipatory activity in adult mice. The functional assembly of this circuit requires innervation by retinal ipRGCs to the SCN during a critical window. The proposed model suggests crosstalk between an FEO (or FEOs) and the IGL, in which IGL neurons act as a node of connection between the FEOs and the central pacemaker in the SCN. Under TRF, inhibitory signals from IGL^{NPY} neurons modulate the SCN function, causing reduced firing activity, and therefore allowing signals from the FEO (or FEOs) to drive robust food-anticipatory activity. The IGL could also be part of the FEO (or FEOs), as previously suggested⁷. IGL^{NPY} neurons send projections to several brain regions and, therefore, a role of any of these non-SCN projections in modulating food-anticipatory activity should not be excluded. How feeding-related stimuli modulate the FEO (or FEOs) and possibly the IGL are unknown. Different humoral signals, such as ghrelin and insulin, are strong candidates for modulating the FEO (or FEOs) and IGL.

Extended Data Table 1 | Systemic metabolic measurements in control and *Opn4^{DTA}* mice housed under free running conditions or TRF.

	Control			<i>Opn4^{DTA}</i>		
	<i>Food ad libitum</i>		Ctrol-TRF	<i>Food ad libitum</i>		DTA-TRF
	Ctrol-CT2	Ctrol-CT14		DTA-CT2	DTA-CT14	
Ghrelin (ng/mL)	2.25 ± 0.30	2.04 ± 0.16 <i>N.S. vs. Ctrol-CT2</i>	4.78 ± 0.62 **<i>(P=0.0015) vs. Ctrol-CT2</i> **<i>(P=0.0035) vs. Ctrol-CT14</i>	2.89 ± 0.41 <i>N.S. vs. Ctrol-CT2</i>	2.06 ± 0.31 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	2.84 ± 0.44 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. DTA-CT14</i> *<i>(P=0.0157) vs. Ctrol-TRF</i>
Insulin (ng/mL)	0.68 ± 0.09	0.92 ± 0.20 <i>N.S. vs. Ctrol-CT2</i>	0.47 ± 0.08 <i>N.S. vs. Ctrol-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	0.91 ± 0.13 <i>N.S. vs. Ctrol-CT2</i>	0.95 ± 0.20 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	0.51 ± 0.09 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. DTA-CT14</i> <i>N.S. vs. Ctrol-TRF</i>
Leptin (ng/mL)	1.88 ± 0.39	3.35 ± 0.73 <i>N.S. vs. Ctrol-CT2</i>	3.73 ± 0.57 <i>N.S. vs. Ctrol-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	3.37 ± 0.73 <i>N.S. vs. Ctrol-CT2</i>	3.99 ± 1.20 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	4.77 ± 0.96 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. DTA-CT14</i> <i>N.S. vs. Ctrol-TRF</i>
Glucose (ng/mL)	104.0 ± 3.9	106.4 ± 2.6 <i>N.S. vs. Ctrol-CT2</i>	105.3 ± 4.2 <i>N.S. vs. Ctrol-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	125.7 ± 9.5 <i>N.S. vs. Ctrol-CT2</i>	104.4 ± 6.2 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. Ctrol-CT14</i>	96.5 ± 8.8 <i>N.S. vs. DTA-CT2</i> <i>N.S. vs. DTA-CT14</i> <i>N.S. vs. Ctrol-TRF</i>

Blood samples were collected at circadian times (CT) 2 and 14 (inactive and active phase, respectively) from mice housed under free-running conditions (ad libitum access to food) or TRF (21st day, immediately before food access). Total ghrelin, insulin, leptin and glucose levels (ng ml⁻¹) were measured for all samples. Data are mean ± s.e.m. (n = 13 mice for each genotype). **P* < 0.05, ***P* < 0.01, two-tailed Tukey's test. Ctrol: control mice; DTA: *Opn4^{DTA}* mice.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Vital-View software (Mini Mitter), version 14.01f3 32-bits
ClockLab (Actimetrics), version 6.0.53
EchoMRI composition analyzer 100H

Data analysis

Adobe Photoshop (Adobe Systems), Version: 19.1.2 20180302.r.277 2018/03/02: 1160083 x64
ImageJ software (NIH, USA), version 2.0.0-rc-43/1.50e
GraphPad Prism, version 7.0a
G*Power 3 software

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw data that support the findings of this study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Calculation of sample size per experiment was determined, or confirm by post hoc analyses, using a G*Power 3 software
Data exclusions	AAV Injection sites were confirmed post hoc by assessing reporter expression. Mice without correct targeting of tracers and/or vectors were excluded from this study.
Replication	Experiments have been replicated multiple times, using different animal cohorts.
Randomization	Mice were randomly assigned to control (sham) or experimental group.
Blinding	Morphological analysis and score analysis were performed with the experimenter blind to genotype and/or condition.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	rabbit α -RFP (MBL PM005, 1:1000) chicken α -GFP (AbCam Ab13970, 1:2000) mouse IgG1 α -c-Fos (EnCor MCA-2H2, 1:1000) rabbit antibody α -NPY (Peninsula lab T-4070, 1:500)
Validation	Antibodies were used according to manufacturer's instructions (protocol, dilutions, as well as positive controls).

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Male and females adult mice were used in this study. Wild-type (WT) mice of a mixed background (B6/129 F1 hybrid, Stock #101043), NPYtm1Rpa (Stock #4545), and NPYcre (Stock #27851) mice were obtained from the Jackson Laboratory. Opn4DTA/DTA, Opn4attnDTA/attnDTA, and Opn4Cre/+;Brn3bDTA/+ mouse lines were previously described
Wild animals	The study did not involve wild animals
Field-collected samples	The study did not involve samples collected from the field
Ethics oversight	All animals were handled in accordance with guidelines of the Animal Care and Use Committees of the National Institute of Mental Health (NIMH).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Ligand-induced monoubiquitination of BIK1 regulates plant immunity

<https://doi.org/10.1038/s41586-020-2210-3>

Received: 12 September 2018

Accepted: 21 February 2020

Published online: 22 April 2020

 Check for updates

Xiyu Ma^{1,2}, Lucas A. N. Claus^{3,4}, Michelle E. Leslie^{5,10,11}, Kai Tao^{6,11}, Zhiping Wu^{7,8,11}, Jun Liu^{1,2}, Xiao Yu^{2,9}, Bo Li^{2,9}, Jinggeng Zhou^{1,2}, Daniel V. Savatin^{3,4}, Junmin Peng^{7,8}, Brett M. Tyler⁶, Antje Heese⁵, Eugenia Russinova^{3,4}, Ping He^{1,2}✉ & Libo Shan^{2,9}✉

Recognition of microbe-associated molecular patterns (MAMPs) by pattern recognition receptors (PRRs) triggers the first line of inducible defence against invading pathogens^{1–3}. Receptor-like cytoplasmic kinases (RLCKs) are convergent regulators that associate with multiple PRRs in plants⁴. The mechanisms that underlie the activation of RLCKs are unclear. Here we show that when MAMPs are detected, the RLCK *BOTRYTIS*-INDUCED KINASE 1 (BIK1) is monoubiquitinated following phosphorylation, then released from the flagellin receptor FLAGELLIN SENSING 2 (FLS2)–BRASSINOSTEROID INSENSITIVE 1-ASSOCIATED KINASE 1 (BAK1) complex, and internalized dynamically into endocytic compartments. The *Arabidopsis* E3 ubiquitin ligases RING-H2 FINGER A3A (RHA3A) and RHA3B mediate the monoubiquitination of BIK1, which is essential for the subsequent release of BIK1 from the FLS2–BAK1 complex and activation of immune signalling. Ligand-induced monoubiquitination and endosomal puncta of BIK1 exhibit spatial and temporal dynamics that are distinct from those of the PRR FLS2. Our study reveals the intertwined regulation of PRR–RLCK complex activation by protein phosphorylation and ubiquitination, and shows that ligand-induced monoubiquitination contributes to the release of BIK1 family RLCKs from the PRR complex and activation of PRR signalling.

Prompt activation of PRRs upon microbial infection is essential for hosts to defend against pathogen attacks^{1–3}. The *Arabidopsis* BIK1 family of RLCKs are immune regulators associated with multiple PRRs, including the bacterial flagellin receptor FLS2 and the BAK1 and SERK family co-receptors^{5,6}. Upon ligand perception, BIK1 is phosphorylated by BAK1 and subsequently dissociates from the FLS2–BAK1 complex⁷. Downstream of the PRR complex, BIK1 phosphorylates plasma-membrane-resident NADPH oxidases to regulate the production of reactive oxygen species (ROS)^{8,9}, and phosphorylates the cyclic nucleotide-gated channels to trigger a rise in cytosolic calcium¹⁰. However, it remains unclear how the activation of BIK1 and its dynamic association with the PRR complex is regulated.

Ligand-induced increase in BIK1 puncta

BIK1–GFP localized both to the periphery of epidermal pavement cells and to intracellular puncta in *Arabidopsis* transgenic plants expressing functional *35S::BIK1-GFP* analysed by spinning disc confocal microscopy (SDCM) (Fig. 1a, Extended Data Fig. 1a, b). BIK1–GFP colocalized with the FM4-64-stained plasma membrane (Fig. 1b), and frequently

within endosomal compartments (Fig. 1b). Time-lapse SDCM showed that BIK1–GFP puncta were highly mobile, disappearing, appearing, and moving rapidly in and out of the plane of view (Extended Data Fig. 1c). The abundance of BIK1–GFP puncta increased over time (3–17 and 18–32 min) after treatment with the flagellin peptide flg22 (Fig. 1c, Extended Data Fig. 1d–p). The timing of the ligand-induced increase in BIK1–GFP puncta differed from that of the increase in FLS2–GFP puncta, which were significantly increased 35 min after flg22 treatment^{11–13} (Fig. 1d). Ligand-induced endocytosis of FLS2 contributes to the degradation of the activated FLS2 receptor and attenuation of signalling^{11–14}, whereas increased abundance of BIK1–GFP puncta precedes that of FLS2–GFP (Fig. 1c, d).

Ligand-induced BIK1 monoubiquitination

Ligand-induced FLS2 degradation is mediated by the U-box E3 ligases PUB12 and PUB13, which polyubiquitinate FLS2^{15–17}. We tested whether BIK1 is ubiquitinated upon treatment with flg22 using an in vivo ubiquitination assay in *Arabidopsis* protoplasts that co-expressed FLAG epitope-tagged ubiquitin (FLAG–UBQ) and haemagglutinin (HA)

¹Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, USA. ²Institute for Plant Genomics and Biotechnology, Texas A&M University, College Station, TX, USA.

³Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. ⁴Center for Plant Systems Biology, VIB, Ghent, Belgium. ⁵Department of Biochemistry, Interdisciplinary Plant Group, University of Missouri-Columbia, Columbia, MO, USA. ⁶Center for Genome Research and Biocomputing and Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA. ⁷Department of Structural Biology, Center for Proteomics and Metabolomics, St Jude Children's Research Hospital, Memphis, TN, USA. ⁸Department of Developmental Neurobiology, Center for Proteomics and Metabolomics, St Jude Children's Research Hospital, Memphis, TN, USA. ⁹Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, USA. ¹⁰Present address: Elemental Enzymes, St Louis, MO, USA. ¹¹These authors contributed equally: Michelle E. Leslie, Kai Tao, Zhiping Wu.

✉e-mail: pinghe@tamu.edu; lshan@tamu.edu

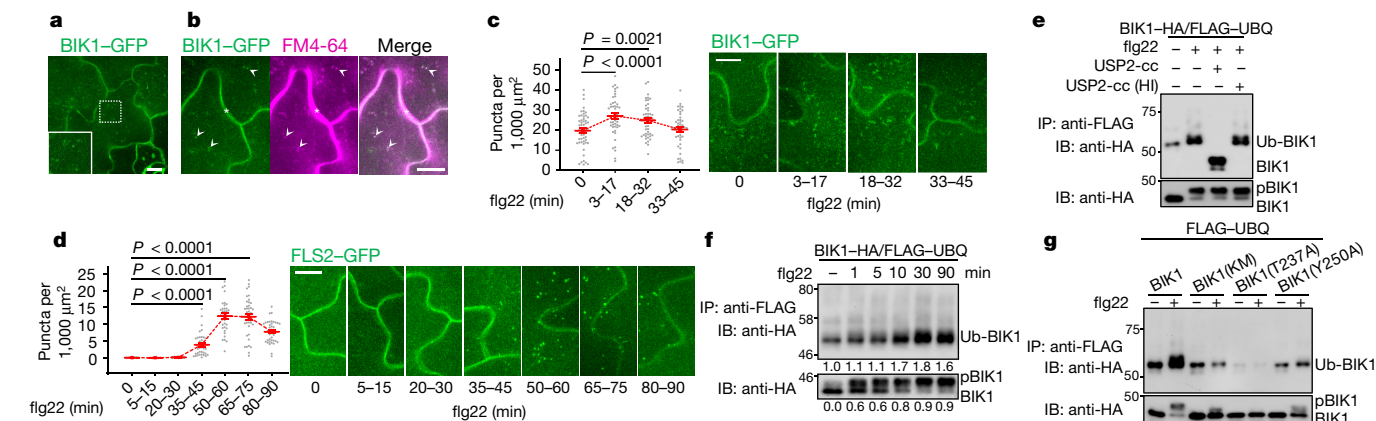


Fig. 1 | MAMP-induced BIK1 endocytosis and monoubiquitination.

a, BIK1-GFP localizes to the cell periphery and intracellular puncta in maximum intensity projections of cotyledon epidermal cells (dashed box expanded in insert). Scale bar, 10 μ m. **b**, BIK1-GFP colocalizes with FM4-64 in the plasma membrane (asterisk) and intracellular puncta (arrowheads). Scale bar, 5 μ m. Pearson's correlation coefficient for BIK1-GFP and FM4-64 is 0.55 ± 0.14 ($n = 35$). **c, d**, BIK1 and FLS2 puncta increase after treatment with 1 μ M flg22. Mean \pm s.e.m. overlaid on dot plots. $n = 56, 48, 49, 47$ images for 0, 3–17, 18–32, 33–45 min of treatment, respectively, for BIK1-GFP (**c**) and $n = 24, 15, 21, 36, 34, 39, 39$ images for 0, 5–15, 20–30, 35–45, 50–60, 65–75, 80–90 min of treatment, respectively, for FLS2-GFP (**d**). Scale bar, 5 μ m (one-way analysis of variance (ANOVA)). **e**, Flg22 induces BIK1 monoubiquitination. Protoplasts from wild-type plants were transfected with plasmids expressing *BIK1-HA* and *FLAG-UBQ*, and were treated with 100 nM flg22 for 30 min. After immunoprecipitation (IP) with anti-FLAG agarose, ubiquitinated BIK1 was

detected by immunoblot (IB) using anti-HA antibodies (lanes 1 and 2) or treated with GST-USP2-cc (lane 3). Heat-inactivated (HI) USP2-cc was used as a control (lane 4). Bottom panel shows BIK1-HA protein expression. Numbers on left show molecular mass (kDa). **f**, Time-course of flg22-induced BIK1 phosphorylation and ubiquitination. Protoplasts expressing FLAG-UBQ and BIK1-HA were treated with 100 nM flg22 for the indicated times. BIK1 band intensities were quantified using Image Lab (Bio-Rad). Quantification of BIK1 phosphorylation (under bottom panel) calculated as ratio of intensity of the upper band (pBIK1) to the sum intensities of shifted and non-shifted bands (pBIK1 + BIK1). Quantification of BIK1 ubiquitination (under top panel) calculated as relative intensity (fold change) of Ub-BIK1 bands (no treatment set to 1.0). **g**, BIK1 variants with impaired phosphorylation show compromised flg22-induced ubiquitination. All experiments were repeated at least three times with similar results.

epitope-tagged BIK1 (Fig. 1e, Extended Data Fig. 2a). Treatment with flg22 induced ubiquitination of BIK1 (Fig. 1e), as ubiquitinated BIK1 was detected by an anti-HA immunoblot upon immunoprecipitation with an anti-FLAG antibody. Flg22 also induced ubiquitination of BIK1 in *pBIK1::BIK1-HA* transgenic plants (Extended Data Fig. 2b). The strong and discrete band of ubiquitinated BIK1 indicates monoubiquitination (Fig. 1e, Extended Data Fig. 2a, b), in contrast to the ladder-like smear of protein migration that indicates polyubiquitination of BAK1 and FLS2 (Extended Data Fig. 2c, d). The apparent molecular mass of ubiquitinated BIK1 (about 52 kDa) is around 8 kDa larger than that of unmodified BIK1 (44 kDa), consistent with the attachment of a single ubiquitin to BIK1. Incubation with the catalytic domain of the mouse deubiquitinase USP2 (USP2-cc), but not its heat-inactivated form, reduced the molecular mass by about 8 kDa (Fig. 1e). We observed a similar pattern of ubiquitination of BIK1 when we used the UBQ(K0) variant, in which all seven lysine residues in UBQ were changed to arginine, thus preventing the formation of polyubiquitination chains (Extended Data Fig. 2e, f). Notably, flg22-induced ubiquitination of BIK1 was blocked by treatment with the ubiquitination inhibitor PYR-41, but not by the proteasome inhibitor MG132, and was not observed in *fls2* or *bak1-4* mutants (Extended Data Fig. 2g–i). In addition to flg22, other MAMPs—including elf18, pep1, and chitin—also induced monoubiquitination of BIK1 (Extended Data Fig. 2j), in line with the notion that BIK1 is a convergent component downstream of multiple PRRs⁴. Monoubiquitination of the BIK1 family RLCKs PBL1 and PBL10, but not of another RLCK, BSK1, was enhanced upon treatment with flg22 (Extended Data Fig. 2k, l), suggesting that detection of MAMPs induces monoubiquitination of BIK1 family RLCKs.

Upon flg22 perception, BIK1 is phosphorylated^{5,6}, as shown by an immunoblot mobility shift within 1 min with a plateau around 10 min (Fig. 1f). However, flg22-induced ubiquitination of BIK1 becomes apparent only 10 min after treatment and reaches a plateau around 30 min (Fig. 1f), suggesting that flg22-induced ubiquitination of BIK1 may occur

after its phosphorylation. BIK1 phosphorylation-deficient mutants, including a kinase-inactive mutant (BIK1(KM)) and two phosphorylation site mutants (BIK1(T237A) and BIK1(Y250A)) showed largely compromised flg22-induced ubiquitination (Fig. 1g). In addition, the kinase inhibitor K252a blocked flg22-induced ubiquitination of BIK1 (Extended Data Fig. 3a). Plasma membrane localization is required for BIK1 ubiquitination, as BIK1(G2A), which bears a mutation of the myristoylation motif that is essential for plasma membrane localization, was not ubiquitinated upon flg22 treatment (Extended Data Fig. 3b, c). Together, these data suggest that flg22-induced phosphorylation of BIK1 is a prerequisite for its monoubiquitination at the plasma membrane.

BIK1 ubiquitination by RHA3A and RHA3B

There are 30 lysine residues in BIK1, each of which could potentially be ubiquitinated. We individually mutated 28 lysine residues to arginine (except for K105 and K106, which are located in the ATP-binding pocket and are required for kinase activity), and screened the mutants for flg22-induced ubiquitination. None of the individual K-to-R mutants blocked the ubiquitination of BIK1 without altering its kinase activity (Extended Data Fig. 3d). BIK1(K204R), in which flg22-induced BIK1 monoubiquitination was compromised, also showed reduced phosphorylation in vivo and in vitro (Extended Data Fig. 3d, e). To identify BIK1-associated regulators, we carried out a yeast two-hybrid screen using BIK1(G2A) as bait, and identified *RHA3A* (*AT2G17450*), which encodes a functionally uncharacterized E3 ubiquitin ligase with a RING-H2 finger domain and an N-terminal transmembrane domain (Fig. 2a). We confirmed that BIK1 interacts with RHA3A using an in vitro pull-down assay (Fig. 2b), an in vivo co-immunoprecipitation (co-IP) assay in *Arabidopsis* protoplasts (Extended Data Fig. 4a), and co-IP in transgenic plants that expressed both *BIK1* and *RHA3A* under their native promoters (Fig. 2c, Extended Data Fig. 4b). RHA3B (which is encoded

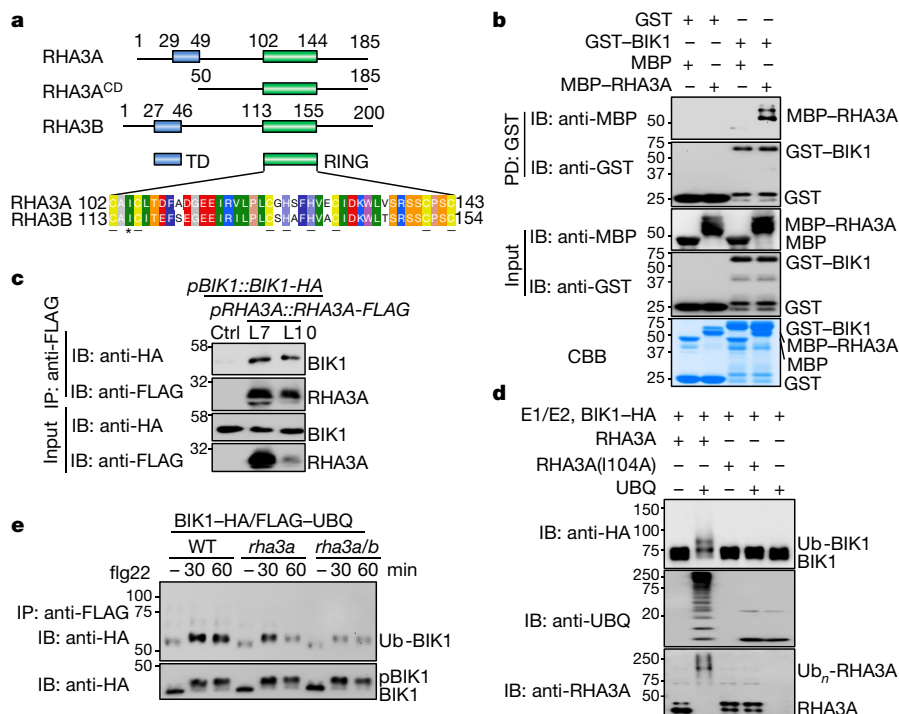


Fig. 2 | The E3 ligases RHA3A/B interact with and monoubiquitinate BIK1. **a**, Domain organization of RHA3A/B. TD, transmembrane domain; RING, E3 catalytic domain; RHA3A^{CD}, cytoplasmic domain. Amino-acid positions and the sequence of RING domain are shown. Cysteine and histidine residues that coordinate zinc are underlined. Asterisk shows the isoleucine residue that is involved in the E2–RING interaction. **b**, BIK1 interacts with RHA3A. GST or GST–BIK1 proteins immobilized on glutathione sepharose beads were incubated with maltose-binding protein (MBP) or MBP–RHA3A^{CD}–HA proteins. Washed beads were subjected to immunoblotting with anti-MBP or anti-GST (top two panels). Input proteins are shown by immunoblotting (middle two panels) and Coomassie blue (CBB) staining (bottom). **c**, BIK1 associates with RHA3A. Transgenic plants carrying *pBIK1::BIK1-HA* and *pRHA3A::RHA3A-FLAG* (lines 7 and 10) were used for co-IP assay with anti-FLAG agarose and immunoprecipitated proteins were immunoblotted with anti-HA or anti-FLAG (top two panels). Bottom two panels, expression of BIK1–HA and RHA3A–FLAG. **d**, RHA3A ubiquitinates BIK1. GST–RHA3A^{CD} or its I104A mutant was used in a ubiquitination reaction containing GST–BIK1–HA, E1, E2, and ATP. **e**, RHA3A/B are required for ubiquitination of BIK1. *rha3a/b* and *rha3a* plants were used for protoplast isolation followed by transfection with plasmids expressing *BIK1-HA* and *FLAG-UBQ*. The experiments were repeated three times with similar results.

by *AT4G35480*) is the closest homologue of RHA3A, bearing 66% amino acid identity (Fig. 2a); RHA3B also co-immunoprecipitated with BIK1 (Extended Data Fig. 4c). Flg22 treatment did not affect the interaction between BIK1 and RHA3A or RHA3B (called RHA3A/B henceforth) (Extended Data Fig. 4a, c). Moreover, RHA3A/B co-immunoprecipitated with FLS2 (Extended Data Fig. 4d).

An in vitro ubiquitination assay showed that RHA3A had autoubiquitination activity and monoubiquitinated itself (Extended Data Fig. 5a, b). Notably, glutathione-S-transferase (GST)–RHA3A, but not GST–RHA3A(I104A), in which a conserved isoleucine residue had been substituted, monoubiquitinated GST–BIK1–HA, as shown on immunoblots by an additional discrete band that migrated with an approximately 8-kDa increase in molecular mass (Fig. 2d). The available *rha3a* and *rha3b* transfer DNA (T-DNA) insertion lines did not show a significant reduction in expression of the corresponding transcripts (Extended Data Fig. 5c). We therefore generated artificial microRNAs (amiRNAs) of *RHA3A/B*¹⁸. Co-expression of *amiR-RHA3A* and *amiR-RHA3B*, but not of *amiR-RHA3A* alone, suppressed flg22-induced monoubiquitination of BIK1 in protoplast transient assays (Extended Data Fig. 5d, e). Flg22-induced BIK1 monoubiquitination, but not phosphorylation, was also reduced in transgenic plants expressing *amiR-RHA3A* and *amiR-RHA3B* driven by the native promoters (Extended Data Fig. 5f, g). We also generated *rha3a* and *rha3a/b* mutants using the CRISPR–Cas9 system (Extended Data Fig. 5h). Flg22-induced monoubiquitination of BIK1 was reduced in the *rha3a/b* mutant (Fig. 2e). These data indicate that RHA3A/B modulate flg22-induced monoubiquitination of BIK1.

Sites of RHA3A-mediated BIK1 ubiquitination

To identify sites of RHA3A-mediated BIK1 ubiquitination, we performed liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis of in vitro ubiquitinated BIK1. Among ten lysine residues identified (Fig. 3a, b, Extended Data Fig. 6a–i), K106 (which resides in the ATP-binding pocket) blocked BIK1 kinase activity when mutated⁷. Among the other nine lysine sites, all six lysines (K95, K170, K186, K286, K337, and K358) for which structural information is available¹⁹ are located

on the surface of BIK1 (Fig. 3c). Furthermore, six ubiquitinated lysine residues were detected by LC–MS/MS of in vivo ubiquitinated BIK1–GFP upon treatment with flg22, and they all overlapped with those detected during in vitro RHA3A–BIK1 ubiquitination reactions (Extended Data Fig. 7a–h). Individual lysine mutations did not affect ubiquitination of BIK1 in vivo (Extended Data Fig. 3d), whereas combined mutations of the N-terminal five lysines (BIK1(N5KR)) or C-terminal four lysines (BIK1(C4KR)) partially compromised flg22-induced BIK1 ubiquitination. Mutation of all nine lysines in BIK1(9KR) largely blocked flg22-induced BIK1 monoubiquitination in vivo (Fig. 3d) and RHA3A-mediated in vitro ubiquitination (Fig. 3e). BIK1(9KR) showed similar activities to BIK1 with regard to its in vitro kinase activity (Fig. 3f), flg22-induced BIK1 phosphorylation, and association with RHA3A in protoplasts (Extended Data Fig. 8a, b). Furthermore, *35S::BIK1^{9KR}-HA/WT* transgenic plants showed normal flg22-induced MAPK activation and ROS production (Extended Data Fig. 8c, d). Collectively, the data indicate that RHA3A monoubiquitinates BIK1 and that phosphorylation of BIK1 does not require monoubiquitination. Notably, BIK1 monoubiquitination may not be restricted to a single lysine, and multiple lysine residues could serve as monoubiquitin conjugation sites. Alternatively, monoubiquitination might be the primary form of modification of BIK1, whereas polyubiquitinated BIK1 could be short-lived.

BIK1 monoubiquitination in immunity

BIK1(9KR), in which monoubiquitination but not phosphorylation of BIK1 is blocked, enabled us to examine the function of BIK1 monoubiquitination without compromised kinase activity. We generated *BIK1^{9KR}* transgenic plants driven by the *BIK1* native promoter in a *bik1* background (*pBIK1::BIK1^{9KR}-HA/bik1*) (Extended Data Fig. 8e, f). Unlike *pBIK1::BIK1-HA/bik1* transgenic plants, *pBIK1::BIK1^{9KR}-HA/bik1* transgenic plants exhibited a reduced flg22-triggered ROS burst similar to that of the *bik1* mutant (Fig. 4a). Moreover, *pBIK1::BIK1^{9KR}-HA/bik1* transgenic plants were more susceptible to the bacterial pathogen *Pseudomonas syringae* pv. *tomato* (*Pst*) DC3000 *hrcC* than were wild-type or *pBIK1::BIK1-HA/bik1* transgenic plants (Fig. 4b). In addition, *amiR-RHA3A/B* transgenic plants exhibited compromised

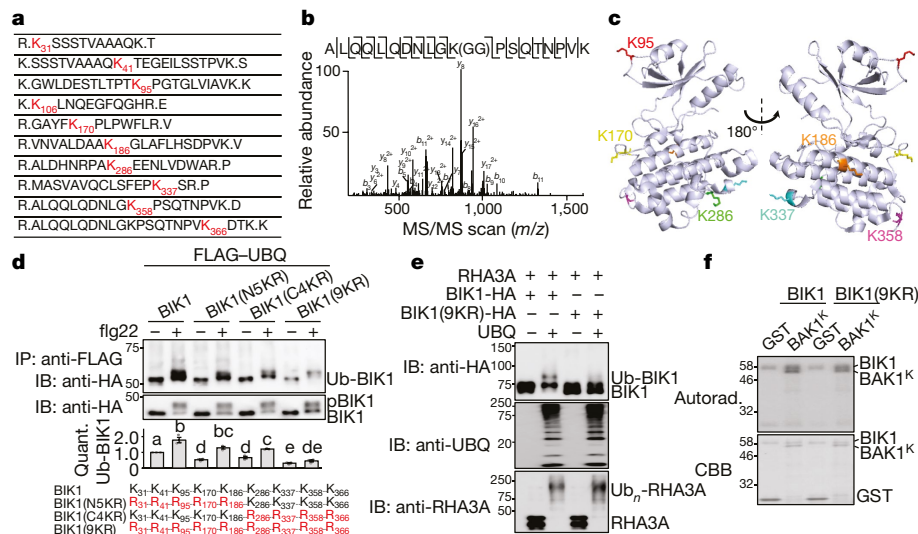


Fig. 3 | Identification of sites of RHA3A-mediated BIK1 ubiquitination.

a, BIK1 is ubiquitinated by RHA3A at multiple lysine residues. Ubiquitinated lysine residues with a diglycine remnant identified by LC-MS/MS analysis are shown in red with amino-acid positions. **b**, MS/MS spectrum of the peptide containing K₃₅₈. **c**, Structure of BIK1 with six lysines identified as ubiquitination sites shown. Structural information was obtained from the Protein Data Bank (PDB ID: 5TOS) and analysed by PyMOL. **d**, BIK1(9KR) shows compromised flg22-induced ubiquitination. FLAG-UBQ and HA-tagged BIK1 mutants were expressed in protoplasts followed by treatment with 100 nM flg22 for 30 min. Quantification of

fold change in BIK1 ubiquitination is shown mean \pm s.e.m. overlaid on dot plot (middle). Different letters indicate significant difference with others (for example, the rightmost bar is significantly different from those marked a, b, and c but not d or e) ($P < 0.05$, one-way ANOVA, $n = 3$). Lysines mutated in BIK1 mutants are shown in red (bottom). **e**, RHA3A cannot ubiquitinate BIK1(9KR). The assay was performed as in Fig. 2d. **f**, BIK1(9KR) exhibits normal in vitro kinase activity. The kinase assay was performed using GST-BIK1 or GST-BIK1(9KR) as the kinase and GST or GST-BAK1^K (kinase domain) as the substrate. All experiments except MS analyses were repeated three times with similar results.

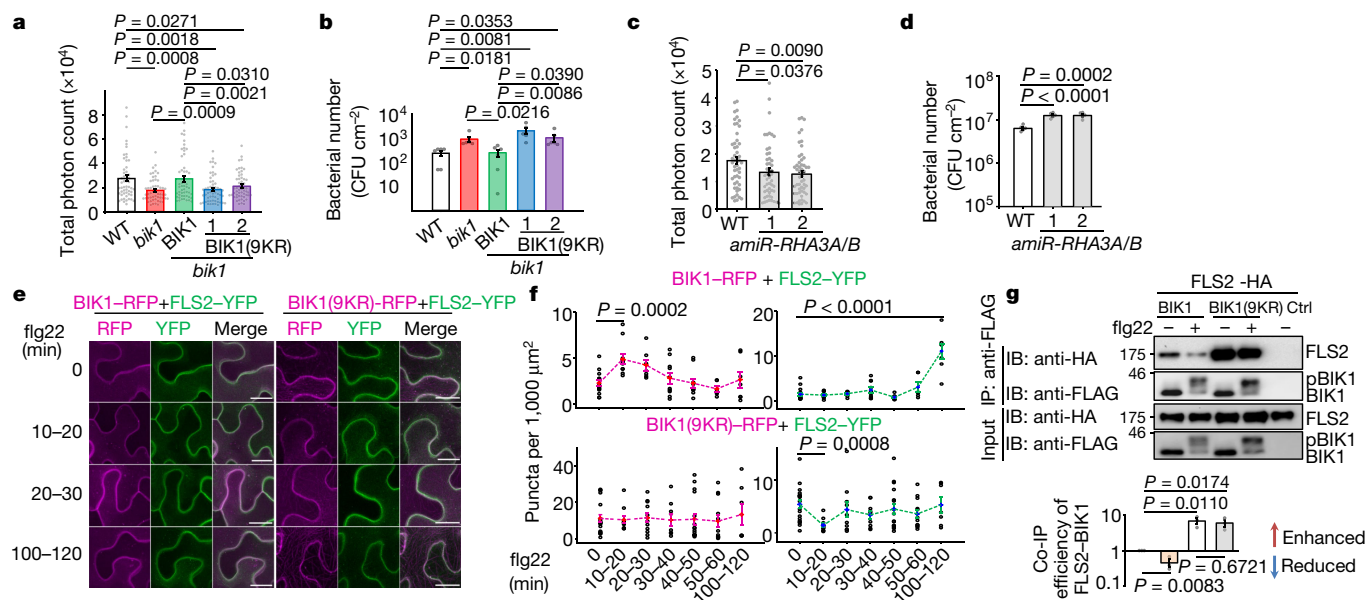


Fig. 4 | RHA3A/B-mediated monoubiquitination of BIK1 contributes to its function in immunity and endocytosis. **a**, *pBIK1::BIK1^{9KR}-HA/bik1* transgenic plants (lines 1 and 2) cannot complement *bik1* for flg22-induced ROS production. One-way ANOVA; wild-type, BIK1/*bik1*: $n = 53$; *bik1*: $n = 54$; BIK1(9KR)/*bik1*: $n = 55$. In all panels, data are shown as mean \pm s.e.m. overlaid on dot plot; lines beneath P values indicate relevant pairwise comparisons. **b**, The *pBIK1::BIK1^{9KR}-HA/bik1* transgenic plants show increased bacterial growth of *Pst* DC3000 *hrcC*⁻. Plants were spray-inoculated and bacterial growth was measured at four days post-inoculation (dpi). One-way ANOVA, $n = 6$. CFU, colony-forming units. **c**, *amiRNA-RHA3A/B* plants show reduced flg22-induced ROS production. One-way ANOVA, $n = 51$. **d**, *amiRNA-RHA3A/B* plants show increased bacterial growth of *Pst* DC3000. Plants were hand-inoculated and bacterial growth was measured at 2 dpi. One-way ANOVA, $n = 5$. **e**, **f**, Flg22-induced endocytosis of

BIK1, BIK1(9KR), and FLS2 in *N. benthamiana* leaf epidermal cells. **e**, BIK1-TagRFP or BIK1(9KR)-TagRFP was co-expressed with FLS2-YFP followed by treatment with 100 μ M flg22 and then imaged at the indicated time points by confocal microscopy. Scale bars, 20 μ m. **f**, Quantification of BIK1-TagRFP (magenta) and FLS2-YFP (green) puncta. One-way ANOVA, additional images and n values shown in Extended Data Fig. 9c. **g**, BIK1(9KR) does not enable flg22-induced dissociation of BIK1 from FLS2. Top, co-IP was performed using protoplasts expressing FLS2-HA and BIK1-FLAG or BIK1(9KR)-FLAG, followed by treatment with 1 μ M flg22 for 15 min. Bottom, the interaction of BIK1 with FLS2 was quantified as intensity from IP: anti-FLAG, IB: anti-HA divided by intensity from IP: anti-FLAG, IB: anti-FLAG. Mean \pm s.e.m. fold change (BIK1 no treatment = 1.0; one-way ANOVA, $n = 3$). All experiments were repeated three times with similar results.

flg22-triggered production of ROS and enhanced susceptibility to *Pst* DC3000 (Fig. 4c, d) and *Pst* DC3000 *hrcC* and to the fungal pathogen *Botrytis cinerea* (Extended Data Fig. 8g, h). Similar results were obtained with the *rha3a/b* mutants (Extended Data Fig. 8i, j). Together, the data indicate that RHA3A/B-mediated monoubiquitination of BIK1 has a role in regulating ROS production and plant immunity.

BIK1 monoubiquitination in endocytosis

As detection of flg22 moderately increased BIK1–GFP endosomal puncta (Fig. 1c), we tested whether monoubiquitination of BIK1 is involved in flg22-triggered BIK1 endocytosis. Fewer FM4-64-labelled puncta were observed in plants expressing BIK1(9KR)–GFP than in those expressing BIK1–GFP after 10 or 15 min of treatment with flg22 (Extended Data Fig. 9a, b). In addition, we compared the flg22-triggered endocytosis of BIK1–TagRFP and BIK1(9KR)–TagRFP when co-expressed with FLS2–YFP in *Nicotiana benthamiana*. As seen in transgenic plants (Fig. 1c, d), endosomal puncta of BIK1–TagRFP increased at 10–20 min, whereas FLS2–YFP puncta increased only after 60 min of flg22 treatment (Fig. 4e, f, Extended Data Fig. 9c). A large portion (about 90%) of flg22-induced BIK1–TagRFP puncta did not colocalize with FLS2–YFP puncta (Extended Data Fig. 9d), suggesting that BIK1 and FLS2 are not likely to be internalized together. This is consistent with the differing ubiquitination characteristics of BIK1 and FLS2 (monoubiquitination versus polyubiquitination, 10 min versus 1 h). When compared to BIK1, BIK1(9KR)–TagRFP was more abundant in puncta before treatment, but the number of puncta did not increase after flg22 treatment (Fig. 4e, f, Extended Data Fig. 9c), indicating that internalization of BIK1(9KR)–TagRFP does not respond to activation of PRRs. In addition, colocalization of BIK1(9KR)–TagRFP with YFP-tagged ARA6 (a plant-specific Rab GTPase that resides on late endosomes²⁰) was substantially reduced when compared to that of BIK1–TagRFP (Extended Data Fig. 9e, f). Notably, flg22-induced endocytosis of FLS2–YFP was absent in the presence of BIK1(9KR)–TagRFP (Fig. 4e, f). Together, our data support the conclusion that ligand-induced monoubiquitination of BIK1 contributes to its internalization from the plasma membrane. Notably, whereas flg22 treatment induced phosphorylation-dependent dissociation of BIK1 from FLS2^{5,6,21}, this effect was largely absent in the case of BIK1(9KR) (Fig. 4g), consistent with the finding that BIK1(9KR) shows impaired FLS2 internalization (Fig. 4e, f). In addition, we observed an increase in the association between BIK1(9KR) and FLS2 without flg22 treatment (Fig. 4g). Treatment with the ubiquitination inhibitor PYR-41 also blocked flg22-induced dissociation of BIK1 from FLS2 and enhanced BIK1–FLS2 association (Extended Data Fig. 10a). Our data indicate that ligand-induced monoubiquitination of BIK1 has an important role in dissociation of BIK1 from the plasma membrane-localized PRR complex, endocytosis of BIK1 and activation of immune signalling (Extended Data Fig. 10b).

Discussion

The BIK1 family RLCKs are central elements of plant PRR signalling, with many layers of regulation^{4,22}. The stability of BIK1 is crucial for maintaining immune homeostasis. The plant U-box proteins PUB25 and PUB26 polyubiquitinate BIK1 and regulate its stability in the steady state²³. This module regulates the homeostasis of non-activated BIK1 without affecting ligand-activated BIK1²³. We have identified a role of RHA3A/B in monoubiquitinating BIK1 and activating PRR signalling, which is distinct from that of PUB25 and PUB26. The levels of BIK1(9KR) proteins in transgenic plants and protoplasts are similar to those of wild-type BIK1 (Extended Data Fig. 10c, d), suggesting that monoubiquitination of BIK1 may not regulate its stability. The nature of protein ubiquitination, including monoubiquitination and polyubiquitination, dictates the distinct fates of substrates, such as proteasome-mediated

protein degradation, nonproteolytic functions of protein kinase activation, and membrane trafficking²⁴. Ligand-induced polyubiquitination of FLS2 by PUB12 or PUB13 promotes degradation of FLS2, thereby attenuating immune signalling^{15,16}, whereas ligand-induced monoubiquitination of BIK1 triggers dissociation of BIK1 from PRR complexes and activates intracellular signalling. Thus, differential ubiquitination and endocytosis of distinct PRR–RLCK complex components are likely to serve as cues to fine-tune plant immune responses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2210-3>.

- Couto, D. & Zipfel, C. Regulation of pattern recognition receptor signalling in plants. *Nat. Rev. Immunol.* **16**, 537–552 (2016).
- Yu, X., Feng, B., He, P. & Shan, L. From chaos to harmony: responses and signaling upon microbial pattern recognition. *Annu. Rev. Phytopathol.* **55**, 109–137 (2017).
- Spoel, S. H. & Dong, X. How do plants achieve immunity? Defence without specialized immune cells. *Nat. Rev. Immunol.* **12**, 89–100 (2012).
- Liang, X. & Zhou, J. M. Receptor-like cytoplasmic kinases: central players in plant receptor kinase-mediated signaling. *Annu. Rev. Plant Biol.* **69**, 267–299 (2018).
- Lu, D. et al. A receptor-like cytoplasmic kinase, BIK1, associates with a flagellin receptor complex to initiate plant innate immunity. *Proc. Natl Acad. Sci. USA* **107**, 496–501 (2010).
- Zhang, J. et al. Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a *Pseudomonas syringae* effector. *Cell Host Microbe* **7**, 290–301 (2010).
- Lin, W. et al. Tyrosine phosphorylation of protein kinase complex BAK1/BIK1 mediates *Arabidopsis* innate immunity. *Proc. Natl Acad. Sci. USA* **111**, 3632–3637 (2014).
- Li, L. et al. The FLS2-associated kinase BIK1 directly phosphorylates the NADPH oxidase RbohD to control plant immunity. *Cell Host Microbe* **15**, 329–338 (2014).
- Kadota, Y. et al. Direct regulation of the NADPH oxidase RBOHD by the PRR-associated kinase BIK1 during plant immunity. *Mol. Cell* **54**, 43–55 (2014).
- Tian, W. et al. A calmodulin-gated calcium channel links pathogen patterns to plant immunity. *Nature* **572**, 131–135 (2019).
- Smith, J. M. et al. Loss of *Arabidopsis thaliana* Dynamin-Related Protein 2B reveals separation of innate immune signaling pathways. *PLoS Pathog.* **10**, e1004578 (2014).
- Beck, M., Zhou, J., Faulkner, C., MacLean, D. & Robatzek, S. Spatio-temporal cellular dynamics of the *Arabidopsis* flagellin receptor reveal activation status-dependent endosomal sorting. *Plant Cell* **24**, 4205–4219 (2012).
- Robatzek, S., Chinchilla, D. & Bolter, T. Ligand-induced endocytosis of the pattern recognition receptor FLS2 in *Arabidopsis*. *Genes Dev.* **20**, 537–542 (2006).
- Smith, J. M., Salamango, D. J., Leslie, M. E., Collins, C. A. & Heese, A. Sensitivity to Flg22 is modulated by ligand-induced degradation and de novo synthesis of the endogenous flagellin-receptor FLAGELLIN-SENSING2. *Plant Physiol.* **164**, 440–454 (2014).
- Lu, D. et al. Direct ubiquitination of pattern recognition receptor FLS2 attenuates plant innate immunity. *Science* **332**, 1439–1442 (2011).
- Zhou, J. et al. The dominant negative ARM domain uncovers multiple functions of PUB13 in *Arabidopsis* immunity, flowering, and senescence. *J. Exp. Bot.* **66**, 3353–3366 (2015).
- Zhou, J. et al. Regulation of *Arabidopsis* brassinosteroid receptor BRI1 endocytosis and degradation by plant U-box PUB12/PUB13-mediated ubiquitination. *Proc. Natl Acad. Sci. USA* **115**, E1906–E1915 (2018).
- Li, J. F., Zhang, D. & Sheen, J. Epitope-tagged protein-based artificial miRNA screens for optimized gene silencing in plants. *Nat. Protocols* **9**, 939–949 (2014).
- Lal, N. K. et al. The receptor-like cytoplasmic kinase BIK1 localizes to the nucleus and regulates defense hormone expression during plant innate immunity. *Cell Host Microbe* **23**, 485–497.e485 (2018).
- Ueda, T., Yamaguchi, M., Uchimiya, H. & Nakano, A. Ara6, a plant-unique novel type Rab GTPase, functions in the endocytic pathway of *Arabidopsis thaliana*. *EMBO J.* **20**, 4730–4741 (2001).
- Lin, W. et al. Inverse modulation of plant immune and brassinosteroid signaling pathways by the receptor-like cytoplasmic kinase BIK1. *Proc. Natl Acad. Sci. USA* **110**, 12114–12119 (2013).
- Lin, W., Ma, X., Shan, L. & He, P. Big roles of small kinases: the complex functions of receptor-like cytoplasmic kinases in plant immunity and development. *J. Integr. Plant Biol.* **55**, 1188–1197 (2013).
- Wang, J. et al. A regulatory module controlling homeostasis of a plant immune kinase. *Mol. Cell* **69**, 493–504.e496 (2018).
- Zhou, B. & Zeng, L. Conventional and unconventional ubiquitination in plant immunity. *Mol. Plant Pathol.* **18**, 1313–1330 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Article

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Plant materials and growth conditions

A. thaliana accession Col-0 (wild type, WT), mutants *fls2*, *bak1-4*, *bik1*, transgenic *pBIK1::BIK1-HA* in the *bik1* background, and *pFLS2::FLS2-GFP* in the Col-0 background have been described previously^{7,13}. *p35S::BIK-GFP* and *p35S::BIK1^{9KR}-GFP* in the Col-0 background, *pBIK1::BIK1^{9KR}-HA* transgenic plants in the *bik1* background, *p35S::BIK1-HA*, *p35S::BIK1^{9KR}-HA* transgenic plants in the Col-0 background, *pBIK1::BIK1-HA* in the Col-0 background, *pBIK1::BIK1-HA/pRHA3A::RHA3A-FLAG* double transgenic plants in the Col-0 background and *pRHA3A::amiR-RHA3A-pRHA3B::amiR-RHA3B* transgenic plants in the Col-0 background were generated in this study (see below). All *Arabidopsis* plants were grown in soil (Metro Mix 366, Sunshine LP5 or Sunshine LC1, Jolly Gardener C/20 or C/GP) in a growth chamber at 20–23 °C, 50% relative humidity and 75 $\mu\text{E m}^{-2} \text{s}^{-1}$ light with a 12-h light/12-h dark photoperiod for four weeks before pathogen infection assay, protoplast isolation, and ROS assay. For confocal microscopy imaging, seeds were sterilized, maintained for 2 days at 4 °C in the dark, and germinated on vertical half-strength Murashige and Skoog ($\frac{1}{2}\text{MS}$) medium (1% (wt/vol) sucrose) agar plates, pH 5.8, at 22 °C in a 16-h light/8-h dark cycle for 5 days with a light intensity of 75 $\mu\text{E m}^{-2} \text{s}^{-1}$. For FM4-64 staining, whole seedlings were incubated for 15 min in 3 ml of $\frac{1}{2}\text{MS}$ liquid medium containing 2 μM FM4-64 and washed twice by dipping into deionized water before adding the elicitor (flg22, 100 nM). Wild-type tobacco (*N. benthamiana*) plants were grown under 14 h of light and 10 h of darkness at 25 °C.

Statistical analyses

Data for quantification analyses are presented as mean \pm s.e.m. The statistical analyses were performed by Student's *t*-test or one-way ANOVA test. The number of replicates is given in the figure legends.

Plasmid construction and generation of transgenic plants

FLS2, BAK1, BIK1, PBL1, PBL10 or BSK1 tagged with HA, FLAG or GFP in a plant gene expression vector pHBT used for protoplast assays, and FLS2^{CD}, BAK1^{CD}, BAK1^K, PUB13, BIK1, or BIK1(KM) fused with GST or MBP used for *Escherichia coli* fusion protein isolation have been described previously^{5,7}. BIK1 point mutations in a pHBT vector were generated by site-directed mutagenesis with primers listed in Supplementary Table 1 using the pHBT-BIK1-HA construct as the template. *BIK1^{NSKR}* was constructed by sequentially mutating K41, K95, K170 and K186 into arginine on *BIK1^{K31R}*. *BIK1^{C4KR}* was constructed by sequentially mutating K337, K358 and K366 on *BIK1^{K286R}*. *pHBT-BIK1^{NSKR}* and *pHBT-BIK1^{C4KR}* were then digested with XbaI and StuI and ligated together to generate *pHBT-BIK1^{9KR}-HA*. *BIK1^{9KR}* was sub-cloned into *pHBT-FLAG* or *pHBT-GFP* with BamHI and StuI digestion to generate *pHBT-BIK1^{9KR}-FLAG* or *pHBT-BIK1^{9KR}-GFP*. *BIK1^{9KR}* was sub-cloned into the binary vector *pCB302-pBIK1::BIK1-HA* or *pCB302-35S::BIK1-HA* with BamHI and StuI digestion to generate *pCB302-pBIK1::BIK1^{9KR}-HA*, or *pCB302-35S::BIK1^{9KR}-HA*. *BIK1-GFP* or *BIK1^{9KR}-GFP* was sub-cloned into *pCB302* with BamHI and PstI digestion to generate *pCB302-35S::BIK1-GFP* and *pCB302-35S::BIK1^{9KR}-GFP*. *BIK1^{K204R}* or *BIK1^{9KR}* was sub-cloned into a modified GST (*pGEX4T-1*, Pharmacia) vector with BamHI and StuI digestion to generate *pGST-BIK1^{K204R}* or *pGST-BIK1^{9KR}*, respectively. *BIK1-HA* or *BIK1^{9KR}-HA* was further sub-cloned into the *pGST* vector as following: digestion with PstI, blunting end by T4 DNA polymerase, digestion with BamHI and ligation into a BamHI/StuI-digested *pGST* vector to generate *pGST-BIK1-HA* and *pGST-BIK1^{9KR}-HA*.

The *RHA3A* gene (*AT2G17450*) was cloned by PCR amplification from Col-0 complementary (c)DNAs with primers containing BamHI at the

5' end and StuI at the 3' end, followed by BamHI and StuI digestion and ligation into the pHBT vector with an HA or FLAG tag at the C terminus. The *RHA3B* gene (*AT4G35480*) was cloned similarly to *RHA3A* using BamHI and SmaI-containing primers. *pHBT-RHA3A^{104A}* was generated by site-directed mutagenesis with primers listed in Supplementary Table 1. *RHA3A^{CD}* (amino acids 50–186) and *RHA3A^{CD/1104A}* were cloned by PCR amplification from *RHA3A* or *RHA3A^{1104A}*, respectively, using BamHI- and StuI-containing primers. *RHA3A^{CD}* and *RHA3A^{CD/1104A}* were sub-cloned into *pGST* or a modified *pMBP* (pMAL-c2, NEB) vector with BamHI and StuI digestion for isolation of *E. coli* fusion proteins. The promoter of *RHA3A* or *RHA3B* was PCR-amplified from genomic DNAs of Col-0 with primers containing SacI and BamHI, and ligated into pHBT. The fragment of *pRHA3A::RHA3A-FLAG* was digested by SacI and EcoRI, and ligated into *pCAMBIA2300*.

AmiRNA constructs were generated as previously described¹⁸. In brief, amiRNA candidates were designed according to instructions at <http://wmd3.weigelworld.org/cgi-bin/webapp.cgi>. Three candidates were chosen for each gene with *RHA3A* for *amiRNA480*: TTTTGT CAATACACTCCACGG; *amiRNA211*: TCAACGCAGATAAGAGCGCTA; *amiRNA109*: TCAAGTAATCTTGACGGTCGT, and *RHA3B* for *amiRNA444*: TTATGCATATTGCACACTCCG; *amiRNA113*: TAATCTAGAGGACCGA GTCAG; *amiRNA214*: TCTACGCATACGAGAGCGCAT. Primers for cloning amiRNAs were generated according to instructions at <http://wmd3.weigelworld.org/cgi-bin/webapp.cgi>. The cognate fragments were cloned into the pHBT-amiRNA-ICE1 vector¹⁸. *pCB302-pRHA3A::amiRNA-RHA3A-pRHA3B::amiRNA-RHA3B* was constructed as follows: the *RHA3A* promoter was PCR amplified from *pRHA3A::RHA3A-FLAG*, digested with SacI and BamHI and ligated with pHBT-amiR-RHA3A to generate pHBT-pRHA3A::amiR-RHA3A. The *pRHA3A::amiR-RHA3A* fragment was further released by SacI and PstI digestion and ligated into pCB302 vector to generate *pCB302-pRHA3A::amiRNA-RHA3A*. *pHBT-pRHA3B::amiR-RHA3B* was constructed similarly followed by PCR amplification using a primer containing SacI sites at both the 5' and 3' ends, subsequent digestion with SacI and ligation into the *pCB302-pRHA3A::amiRNA-RHA3A* vector. Tandem *pRHA3A/B-amiRNA-RHA3A/B* in the same direction was confirmed by digestion and selected for further experiments.

The *rha3a/b* mutant was generated by the CRISPR–Cas9 system following the published protocol²⁵. In brief, primers containing guide RNA (gRNA) sequences of *RHA3A* and *RHA3B* were used in PCR to insert both gRNA sequences into the *pDTIT2* vector. The *pDTIT2* vector containing both gRNAs was further PCR amplified, digested with BsaI and ligated into a binary vector *pHEE401E*. *Agrobacterium-tumefaciens*-mediated floral dip was used to transform the *pHEE401E* vector into Col-0 plants. Genomic DNAs from hygromycin (25 $\mu\text{g/ml}$)-positive plants were extracted, PCR amplified with gene-specific primers and sequenced by Sanger sequencing.

The monomer ubiquitin of *Arabidopsis* ubiquitin gene 10 (*UBQ10*, *At4g05320*) carrying lysine-to-arginine mutations at all the seven lysine residues (*UBQ^{K0}*: 5'-ATGCAGATCTTTGT TAGGACTCTCACCGGAAGGACTATCACCTCGAGGTGGAAAGCTCTGA CACCATCGACAACGTTAGGGCCAGGATCCAGGATAGGGAAGGTATTCC TCCGGATCAGCAGAGGCTTATCTTCGCCGGAAGGCAGTTGG AGGATGG CCGCAGCTTGGCGGATTACAATATCCAGGGAATCCA CCCTCCACTT GGTCTCAGGCTCCGTGGTGTAA-3') was synthesized and cloned into a *pUC57* vector by GenScript USA Incorporation. *UBQ^{K0}* was then amplified by PCR with primers listed in the Supplementary Table 1 and further sub-cloned into a modified pHBT vector with BamHI and PstI digestion to generate pHBT-FLAG-UBQ^{K0}.

Plasmids used for transient expression in *N. benthamiana* were constructed as reported previously²⁶. In brief, *FLS2*, *BIK1*, and *BIK1^{9KR}* were PCR amplified and recombined into *pDONR207-YFP*, *pDONR207-TagRFP*, and *pDONR207-GFP* vectors by In-Fusion HD Cloning (TaKaRa Bio). The *pDONR207* vectors were subsequently transferred to a destination vector *pmAEV* (derived from binary vector *pCAMBIA*

with a 35S promoter) using the Gateway LR reaction (Thermo Fisher scientific).

DNA fragments cloned into the final constructs were confirmed via Sanger sequencing. *A. tumefaciens*-mediated floral dip was used to transform the above binary vectors into *bik1* or Col-0 plants. The transgenic plants were selected using glufosinate-ammonium (Basta, 50 µg/ml) for the *pCB302* vector or kanamycin (50 µg/ml) for the *pCAMBIA2300* vector. Multiple transgenic lines were analysed by immunoblotting for protein expression. Two lines with 3:1 segregation ratios for antibiotic resistance in the T3 generation were selected to obtain homozygous seeds for further studies. *amiR-RHA3A/B* transgenic plants that were resistant to Basta in the T2 generation were used for assays.

Yeast two-hybrid screen

The cDNA library constructed in a modified pGADT7 vector (Clontech) has been previously described¹⁵. BIK1(G2A) from *pHBT-BIK1^{G2A}-HA* was sub-cloned into a modified *pGBKT7* vector with BamHI and StuI digestion. *pGBK-BIK1^{G2A}* was transformed into the yeast AH109 strain. The resulting yeast transformants were then transformed with the cDNA library and screened in synthetic defined (SD) medium without Trp, Leu, His, Ade (SD-T-L-H-A) and SD-T-L-H containing 1 mM 3-amino-1, 2, 4-triazole (3-AT). The confirmed yeast colonies were subjected to plasmid isolation and sequencing.

Pathogen infection assays

Pst DC3000 was cultured overnight at 28 °C in King's B medium supplemented with rifamycin (50 µg/ml). Bacteria were collected by centrifugation at 3,000g, washed and re-suspended to a density of 10⁶ colony-forming units (cfu)/ml with 10 mM MgCl₂. Leaves from four-week-old plants were hand-inoculated with bacterial suspension using a needleless syringe. To measure in planta bacterial growth, five to six sets of two leaf discs, 6 mm in diameter, were punched and ground in 100 µl ddH₂O. Serial dilutions were plated on TSA plates (1% tryptone, 1% sucrose, 0.1% glutamic acid and 1.8% agar) containing 25 µg/ml rifamycin. Plates were incubated at 28 °C and bacterial cfu were counted 2 days after incubation. For spray inoculation, *Pst* DC3000 or *Pst* DC3000 *hrcC*⁻ bacteria were collected and re-suspended to 5 × 10⁸ cfu/ml with 10 mM MgCl₂, silwet L-77 (0.02%) and sprayed onto the leaf surface. Plants were covered with a transparent plastic dome to maintain humidity after spraying. After incubation, the third pair of true leaves was detached, soaked in 70% ethanol for 30 s and rinsed in water, and bacterial growth was measured as described above.

Protoplast transient expression and co-IP assays

Protoplast isolation and the transient expression assay have been described previously²⁷. For protoplast-based co-IP assays, protoplasts were transfected with a pair of constructs (the empty vectors as controls, 100 µg DNA for 500 µl protoplasts at a density of 2 × 10⁵/ml for each sample) and incubated at room temperature for 6–10 h. After treatment with flg22 at the indicated concentrations and time points, protoplasts were collected by centrifugation and lysed in 300 µl co-IP buffer (150 mM NaCl, 50 mM Tris-HCl, pH 7.5, 5 mM EDTA, 0.5% Triton, 1 × protease inhibitor cocktail, before use, adding 2.5 µl 0.4 M DTT, 2 µl 1 M NaF and 2 µl 1 M Na₃VO₃ for 1 ml IP buffer) by vortexing. After centrifugation at 10,000g for 10 min at 4 °C, 30 µl supernatant was collected for input controls and 7 µl anti-FLAG-agarose beads were added to the remaining supernatant and incubated at 4 °C for 1.5 h. Beads were collected and washed three times with washing buffer (150 mM NaCl, 50 mM Tris-HCl, pH 7.5, 5 mM EDTA, 0.5% Triton) and once with 50 mM Tris-HCl, pH 7.5. Immunoprecipitates were analysed by immunoblotting with the indicated antibodies. The amiRNA candidate screens were performed as previously described¹⁸.

In vivo ubiquitination assay

FLAG-tagged UBQ (FLAG-UBQ) or a vector control (40 µg DNA) was co-transfected with the target gene with an HA tag (40 µg DNA) into 400 µl protoplasts at a density of 2 × 10⁵/ml for each sample, and protoplasts were incubated at room temperature for 6–10 h. After treatment with 100 nM flg22 at the indicated time points, protoplasts were collected for co-IP assay in co-IP buffer containing 1% Triton X-100. PYR-41 (Sigma, cat # N2915) was added at the indicated concentrations and time points (see Figure legends).

Recombinant protein isolation and in vitro kinase assays

Fusion proteins were produced from *E. coli* BL21 at 16 °C using LB medium with 0.25 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). GST fusion proteins were purified with Pierce glutathione agarose (Thermo Scientific), and MBP fusion proteins were purified using amylose resin (New England Biolabs) according to the standard protocol from companies. The in vitro kinase assays were performed with 0.5 µg kinase proteins and 5 µg substrate proteins in 30 µl kinase reaction buffer (10 mM Tris-HCl, pH 7.5, 5 mM MgCl₂, 2.5 mM EDTA, 50 mM NaCl, 0.5 mM DTT, 50 µM ATP and 1 µCi [γ-³²P] ATP). After gentle shaking at room temperature for 2 h, samples were denatured with 4 × SDS loading buffer and separated by 10% SDS-PAGE gel. Phosphorylation was analysed by autoradiography.

In vitro ubiquitination assay

Ubiquitination assays were performed as previously described with modifications²⁸. Reactions containing 1 µg substrate, 1 µg HIS₆-E1 (AtUBA1), 1 µg HIS₆-E2 (AtUBC8), 1 µg GST-E3, 1 µg ubiquitin (Boston Biochem, cat # U-100AT-05M) in the ubiquitination reaction buffer (20 mM Tris-HCl, pH 7.5, 5 mM MgCl₂, 0.5 mM DTT, 2 mM ATP) were incubated at 30 °C for 3 h. The ubiquitinated proteins were detected by immunoblotting with indicated antibodies. The rabbit monoclonal anti-RHA3A antibody was generated according to the company's standard protocol against the peptide: AGGDSPSPNKLKKC (GenScript).

In vitro deubiquitination assay

Mouse USP2-cc was cloned by PCR amplification from mouse cDNAs with primers containing BamHI at the 5' end and SmaI at the 3' end, followed by BamHI and SmaI digestion and ligation into the *pGST* vector to construct *pGST-USp2-cc*. GST-USP2-cc fusion proteins were produced in *E. coli* BL21 and purified with Pierce glutathione agarose (Thermo Scientific) according to the manufacturer's standard protocol. Deubiquitination (DUB) assays were performed as previously described with modifications²⁹. In brief, an in vitro ubiquitination assay was performed overnight at 28 °C as described above. The reaction was aliquoted into individual tubes containing USP2-cc or heat-inactivated (HI) (95 °C for 5 min) USP2-cc as a control in the DUB reaction buffer (50 mM Tris-HCl, pH 7.5, 50 mM NaCl and 5 mM DTT) and incubated at 28 °C for 5 h. Samples were then denatured and analysed by immunoblotting.

For in vitro DUB assay with flg22-induced ubiquitinated BIK1, BIK1-HA and FLAG-UBQ were expressed in *Arabidopsis* protoplasts treated with 100 nM flg22 for 30 min. The ubiquitinated BIK1-HA proteins were immunoprecipitated as described above. After washing with 50 mM Tris-HCl, agarose beads were washed once with DUB dilution buffer (25 mM Tris-HCl, pH 7.5, 150 mM NaCl and 10 mM DTT) and mixed with GST-USP2-cc in DUB reaction buffer. After overnight incubation, beads were denatured in SDS buffer and analysed by immunoblotting.

MAPK assay

Five 11-day-old *Arabidopsis* seedlings per treatment, grown on vertical plates with ½MS medium, were transferred into water overnight before flg22 treatment. Seedlings were collected, drilled and lysed in 100 µl co-IP buffer. Protein samples with 1 × SDS buffer were separated in 10%

Article

SDS–PAGE gel to detect pMPK3, pMPK6 and pMPK4 by immunoblotting with anti-pERK1/2 antibody (Cell Signaling, cat # 9101).

Detection of ROS production

The third or fourth pair of true leaves from 4- to 5-week-old soil-grown *Arabidopsis* plants were punched into leaf discs (diameter 5 mm). Leaf discs were incubated in 100 μ l ddH₂O with gentle shaking overnight. Water was replaced with 100 μ l reaction solution containing 50 μ M luminol, 10 μ g/ml horseradish peroxidase (Sigma-Aldrich) supplemented with or without 100 nM flg22. Luminescence was measured with a luminometer (GloMax-Multi Detection System, Promega) with a setting of 1 min as the interval for 40–60 min. Detected values of ROS production were indicated as means of relative light units (RLU).

In vitro GST pull-down assay

GST or GST–BIK1 agarose beads were obtained after elution and washed with 1 \times PBS (137 mM NaCl, 2.7 mM KCl, 15 mM Na₂HPO₄, 4.4 mM KH₂PO₄) three times. HA-tagged MBP–RHA3A^{CD} or MBP proteins (2 μ g) were pre-incubated with 10 μ l prewashed glutathione agarose beads in 300 μ l pull-down incubation buffer (20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 0.1 mM EDTA, and 0.2% Triton X-100) for 30 min at 4 °C. Five microlitres of GST or GST–BIK1 agarose beads were pre-incubated with 20 μ g bovin serum albumin (BSA, Sigma, cat # A7906) in 300 μ l incubation buffer for 30 min at 4 °C with gentle shaking. The supernatant containing MBP–RHA3A^{CD} or MBP was incubated with pre-incubated GST or GST–BIK1 agarose beads for 1 h at 4 °C with gentle shaking. The agarose beads were precipitated and washed three times in pull-down wash buffer (20 mM Tris-HCl, pH 7.5, 300 mM NaCl, 0.1 mM EDTA, and 0.5% Triton X-100). The pulled-down proteins were analysed by immunoblotting with an anti-MBP antibody (Biolegend, cat # 906901).

Mass spectrometry analysis of ubiquitination sites

In vitro ubiquitination reactions with GST–RHA3A^{CD} and GST–BIK1 or GST–BIK1(K204R) were performed as mentioned above with overnight incubation. Reactions were loaded on an SDS–PAGE gel (7.5%) and ran for a relatively short time until the ubiquitinated bands could be separated from the original GST–BIK1 (GST–BIK1 band ran less than 0.5 cm from the separating gel). Ubiquitinated bands were sliced and trypsin-digested before LC–MS/MS analysis on an LTQ-Orbitrap hybrid mass spectrometer (Thermo Fisher) as previously described³⁰. The MS/MS spectra were analysed with SEQUEST software, and images were exported from SEQUEST.

In vivo BIK1 ubiquitination sites were identified as follows: 20 ml of wild-type *Arabidopsis* protoplasts at a concentration of 2×10^5 per ml were transfected with BIK1–GFP and FLAG–UBQ and the protoplasts were treated with 200 nM flg22 for 30 min after 7 h of incubation. GFP-trap-Agarose beads (Chromotek, cat # gta-20) were incubated with cell lysates at a ratio of 10 μ l beads to 4×10^5 cells for 1 h at 4 °C and beads were pooled from 10 tubes, washed using IP buffer three times, and denatured in SDS buffer. Samples were separated by 10% SDS–PAGE and stained with GelCode Blue Stain Reagent (Thermo Fisher cat # 24590). Ubiquitinated bands were sliced and analysed as described above.

Confocal microscopy and image analysis

For laser scanning confocal microscopy, images were taken using a Leica SP8X inverted confocal microscope equipped with a HC PL APO CS2 40 \times /1.10 and 63 \times /1.20 water-corrected objective. The excitation wavelength was 488 nm for both GFP and FM4-64 (Thermo Fisher T13320), 514 nm for YFP and 555 nm for TagRFP using the white light laser. Emission was detected at 500–530 nm for GFP, 570–670 nm for FM4-64, 519–549 nm for YFP, and 569–635 nm for TagRFP by using Leica hybrid detectors. Autofluorescence was removed by adjusting the time gate window between 0.8 and 6 ns. Intensities were manipulated using ImageJ software.

For SDCM, image series were captured using a custom Olympus IX-71 inverted microscope equipped with a Yokogawa CSU-X1 5,000 rpm spinning disc unit and 60 \times silicon oil objective (Olympus UPlanSApo 60 \times /1.30 Sil) as previously described¹¹. For the custom SDCM system, GFP and FM4-64 were excited with a 488-nm diode laser and fluorescence was collected through a series of Semrock Brightline 488-nm single-edge dichroic beamsplitter and bandpass filters: 500–550 nm for GFP and 590–625 nm for FM4-64. Camera exposure time was set to 150 ms. For each image series, 67 consecutive images at a z-step interval of 0.3 μ m (20 μ m total depth) were captured using Andor iQ2 software (Belfast, UK). Images captured by custom SDCM were processed with the Fiji distribution of ImageJ 1.51 (<https://fiji.sc/>) software, and BIK1–GFP and FLS2–GFP endosomal puncta were quantified as the number of puncta per 1,000 μ m² as previously described^{11,31}, with the exception that puncta were detected within a size distribution of 0.1–2.5 μ m². For colocalization of BIK1–GFP with FM4-64 by custom SDCM, cotyledons were stained with 2.5 μ M FM4-64 for 10 min, washed twice, and imaged after a 5-min chase.

For quantification of flg22-induced puncta containing BIK1–GFP or BIK1(9KR)–GFP over time, the maximum number of FM4-64 labelled spots per image area was set to 100%, and the percentage of GFP-colocalizing spots per time interval relative to the maximum was calculated; 20–25 images per time interval, captured from 5 individual plants per genotype were used for quantification.

For transient expression in *N. benthamiana*, *Agrobacterium* strain C58 carrying the constructs of interest was co-infiltrated in the abaxial side of tobacco leaves as described previously³². Between 48 and 72 h after infiltration, multiple infiltrated leaves were treated with 100 μ M flg22 and imaged at the indicated time points. The number of puncta per 1,000 μ m² was quantified as previously described^{11,31}. The percentage colocalization of BIK1 and FLS2 was calculated by dividing the number of BIK1–FLS2 colocalizing puncta by the total number of BIK1 puncta. The percentage colocalization of BIK1 and ARA6 was calculated by dividing the number of BIK1–ARA6 or BIK1(9KR)–ARA6 colocalizing puncta by the total number of ARA6 puncta.

qRT–PCR analysis

Total RNA was isolated from the leaves of four-week-old plants with TRIzol reagent (Invitrogen). One microgram of total RNA was treated with RNase-free DNase I (New England Biolabs) followed by cDNA synthesis with M-MuLV reverse transcriptase (New England Biolabs) and oligo(dT) primer. qRT–PCR analysis was performed using iTaq SYBR green Supermix (Bio-Rad) with primers listed in Supplementary Table 1 in a Bio-Rad CFX384 Real-Time PCR System. The expression of *RHA3A* and *RHA3B* was normalized to the expression of *ACTIN2*.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data supporting the findings of this study are available within the paper and its Supplementary Information files. Source Data (gels and graphs) for Figs. 1–4 and Extended Data Figs. 1–10 are provided with the paper.

- Wang, Z. P. et al. Egg cell-specific promoter-controlled CRISPR/Cas9 efficiently generates homozygous mutants for multiple target genes in *Arabidopsis* in a single generation. *Genome Biol.* **16**, 144 (2015).
- Tao, K., Waletich, J. R., Arredondo, F. & Tyler, B. M. Manipulating endoplasmic reticulum-plasma membrane tethering in plants through fluorescent protein complementation. *Front. Plant Sci.* **10**, 635 (2019).
- He, P., Shan, L. & Sheen, J. The use of protoplasts to study innate immune responses. *Methods Mol. Biol.* **354**, 1–9 (2007).
- Zhou, J., He, P. & Shan, L. Ubiquitination of plant immune receptors. *Methods Mol. Biol.* **1209**, 219–231 (2014).

29. Hospenthal, M. K., Mevissen, T. E. T. & Komander, D. Deubiquitinase-based analysis of ubiquitin chain architecture using Ubiquitin Chain Restriction (UbiCRest). *Nat. Protocols* **10**, 349–361 (2015).
30. Xu, P. et al. Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. *Cell* **137**, 133–145 (2009).
31. Leslie, M. E. & Heese, A. Quantitative analysis of ligand-induced endocytosis of FLAGELLIN-SENSING 2 using automated image segmentation. *Methods Mol. Biol.* **1578**, 39–54 (2017).
32. Boruc, J. et al. Functional modules in the *Arabidopsis* core cell cycle binary protein-protein interaction network. *Plant Cell* **22**, 1264–1280 (2010).

Acknowledgements We thank Q. Chen for the CRISPR–Cas9 system, J. Li and J. Sheen for the amiRNA system, P. de Figueiredo for mouse cDNA, and T. Devarenne, M. Dickman, C. Kaplan, T. Igumenova, J. Sheen, and the members of the Shan and He laboratories for discussion and comments on this work. The work was supported by NIH (R01GM097247) and the Robert A. Welch foundation (A-1795) to L.S., National Science Foundation (NSF) (MCB-1906060) and NIH (R01GM092893) to P.H., NSF (IOS-1147032) to A.H., the Special Research Fund (BOF15/24J/048) to E.R., and NIH (R01GM114260) to J.P.

Author contributions X.M., P.H. and L.S. conceived and designed the experiments; X.M. performed most of the molecular, biochemical and transgenic experiments; L.A.N.C. and

D.V.S., under the supervision of E.R., conducted BIK1, BIK1(9KR) and FLS2 endocytosis experiments in the transgenic plants and *N. benthamiana*, colocalization with ARA6, and spatial and temporal dynamics of BIK1 and FLS2 endocytosis in *N. benthamiana*; M.E.L., under the supervision of A.H., performed BIK1 and FLS2 endocytosis in transgenic plants by spinning disc confocal microscopy; K.T., under the supervision of B.M.T., conducted BIK1(G2A) localization, BIK1 endocytosis and colocalization with ARA6 experiments in *N. benthamiana*; Z.W., under the supervision of J.P., identified BIK1 in vivo and in vitro ubiquitination sites with LC–MS/MS; J.L. performed co-IP and *Botrytis* infection assays; X.Y. performed co-IP and transgenic plant assays; B.L. generated 35S::BIK1-GFP transgenic plants; J.Z. performed in vitro kinase assays; E.R., A.H., B.T., and J.P. analysed data, provided critical feedback and helped to shape the research. All experiments were independently reproduced in different laboratories. X.M., P.H. and L.S. wrote the manuscript with input from all authors.

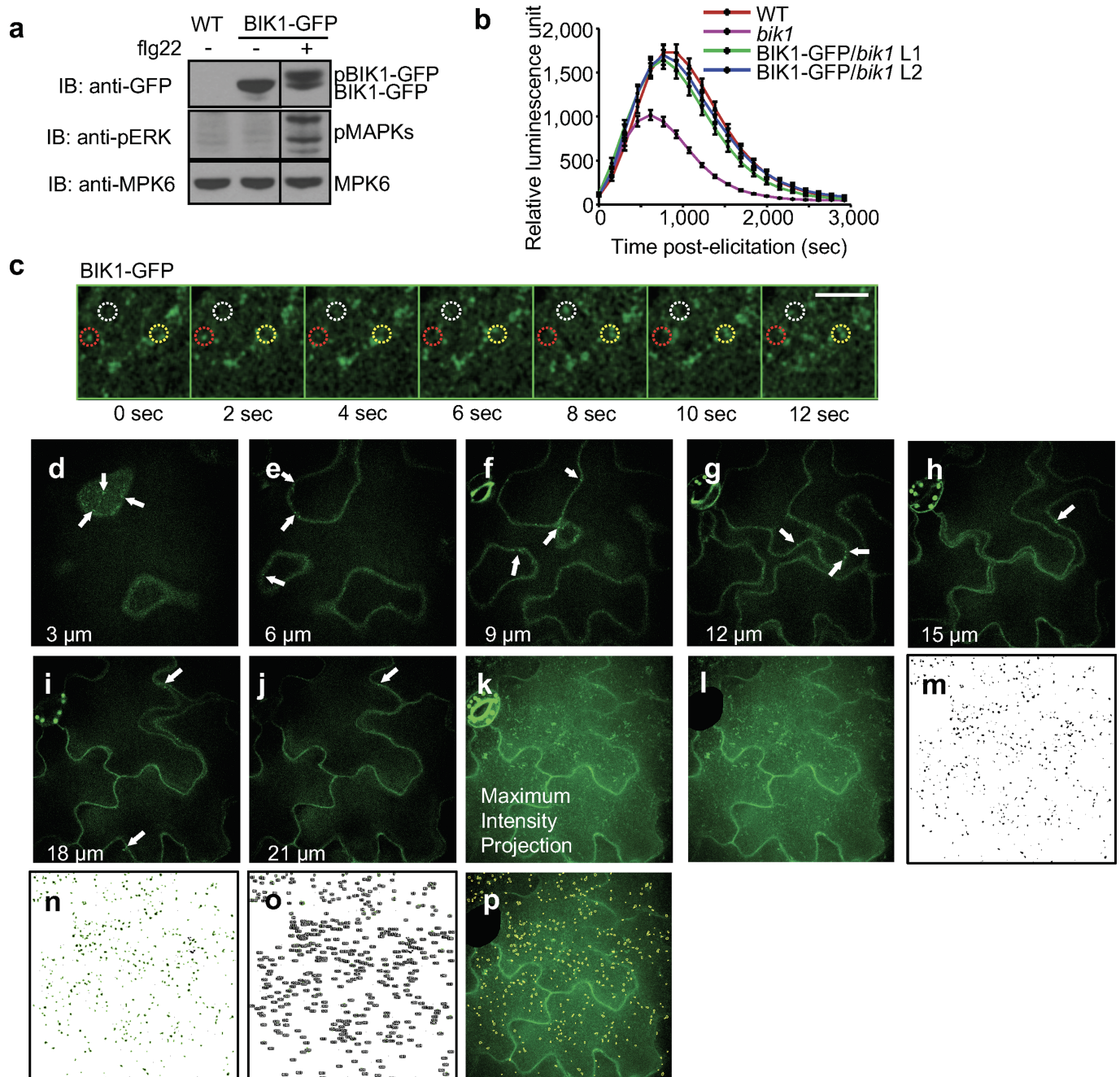
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2210-3>.

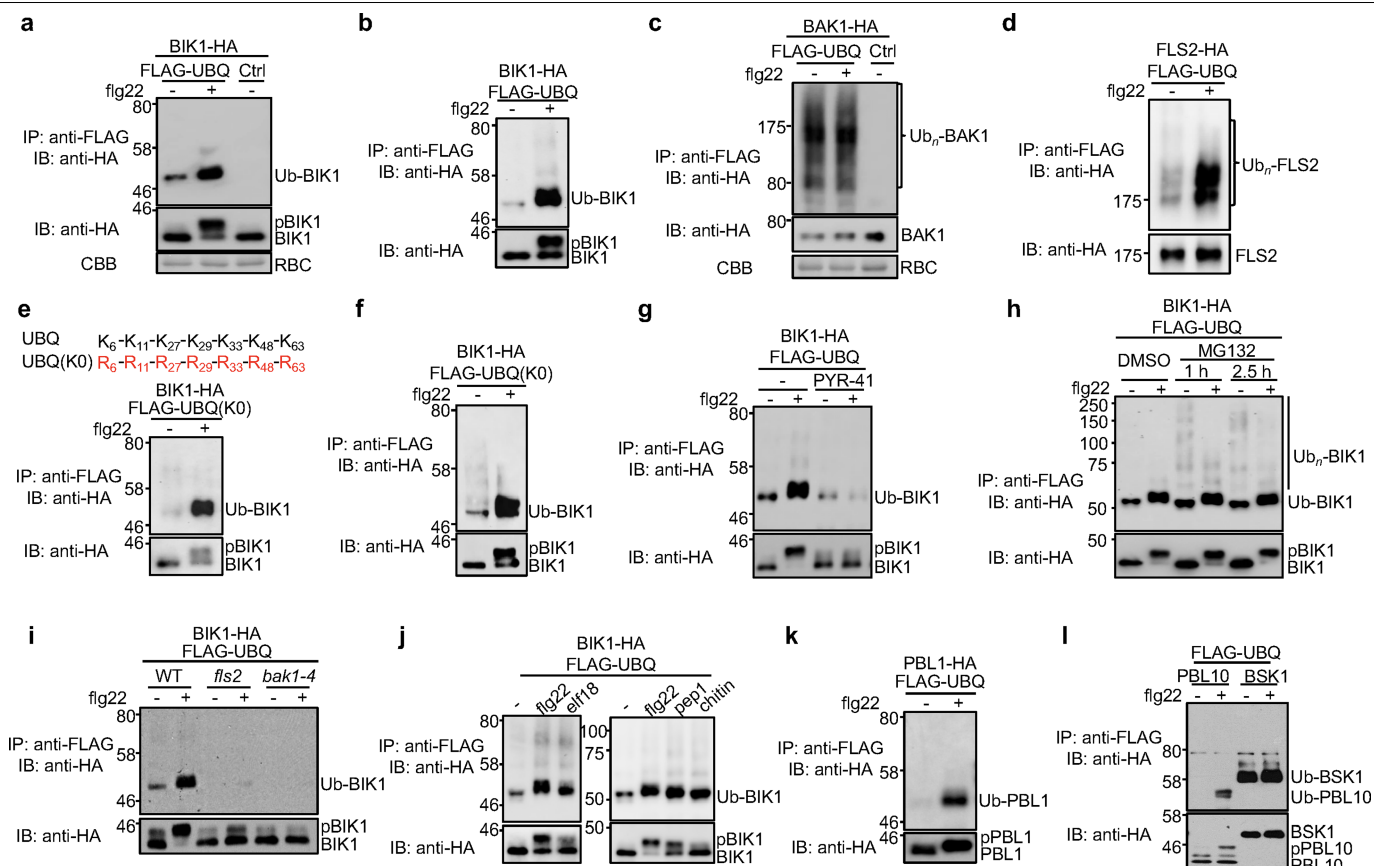
Correspondence and requests for materials should be addressed to P.H. or L.S.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



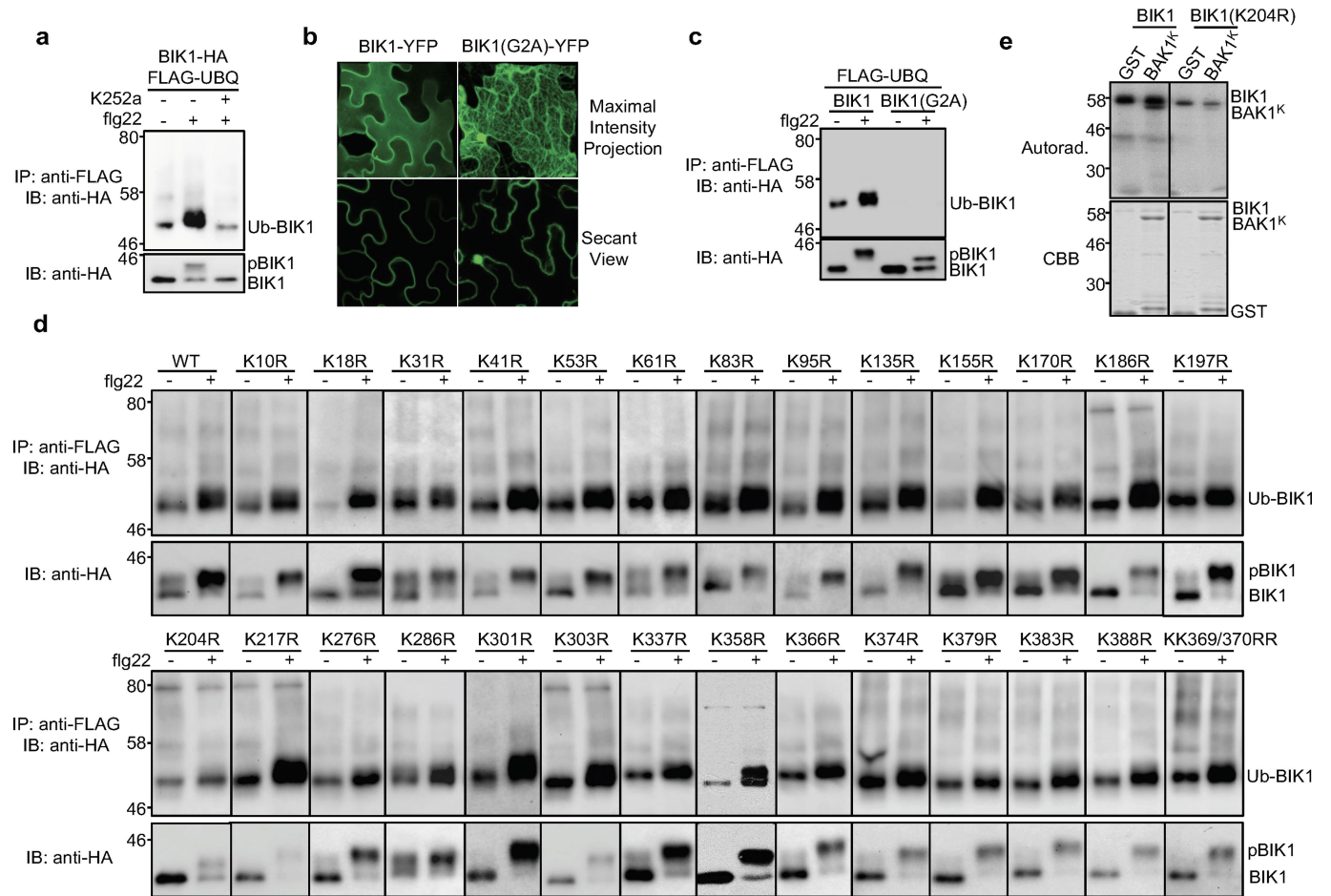
Extended Data Fig. 1 | BIK1-GFP is functional in plants and undergoes endocytosis. **a**, BIK1-GFP is functional, as confirmed by BIK1 phosphorylation in 35S::BIK1-GFP-expressing Col-0 cotyledons after treatment with 1 μ M flg22. MPK6 is a loading control and the black stippled line indicates discontinuous segments from the same gel. **b**, BIK1-GFP restored ROS production in *bik1* leaves upon flg22 treatment. Leaf discs from wild-type, *bik1* and BIK1-GFP complementation plants (lines 1 and 2) were treated with 100 nM flg22 for ROS measurement using a luminometer over 50 min. Data are shown as mean \pm s.e.m. (wild-type, *bik1*: $n = 42$; BIK1-GFP/*bik1*: $n = 45$). **c**, Time-lapse SDCM shows that BIK1-GFP endosomal puncta are highly mobile with puncta that disappear (red circle), appear (yellow circle), and rapidly move in and out of the plane of view (white circle). Scale bar, 5 μ m. **d-k**, BIK1-GFP localizes to endosomal puncta and plasma membrane in cross-sectional images of epidermal cells. The abaxial epidermal cells of cotyledons expressing BIK1-GFP were imaged with SDCM with a Z-step of 0.3 μ m. A subset of the cross-sectional images is shown at the indicated depths (3, 6, 9, 12, 15, 18 and 21 μ m) along with the maximum-intensity projection (MIP) of all 67 images

through the epidermis. BIK1-GFP localizes to both plasma membrane and endosomal puncta (white arrows) within all sections. **k-p**, Method for quantification of BIK1-GFP puncta within MIPs of SDCM images. **k**, MIPs were generated using Fiji distribution of ImageJ 1.51 (<https://fiji.sc/>) for each Z-series captured by SDCM imaging of BIK1-GFP cotyledons. **l**, Regions of MIP with non-pavement cells (for example, stomata) were removed from the image using the line draw and crop functions. The total surface area (μm^2) of the image was measured using the analyze measure function. **m**, Puncta within the cropped MIP were recognized using a customized model generated and applied with the Trainable Weka Segmentation plug-in for Fiji. The same model was applied to all images to generate binary images showing the physical locations of all BIK1-GFP puncta (black). **n-o**, Puncta within the size range 0.1–2.5 μm^2 were highlighted in green (**n**) and counted (**o**) using the analyze particles function in Fiji. BIK1-GFP endocytosis was quantified as the number of puncta per 1,000 μm^2 . **p**, An overlay of the BIK1-GFP puncta (yellow highlight) over the cropped MIP confirmed correct identification of puncta. The experiments in **a-c** were repeated three times with similar results.



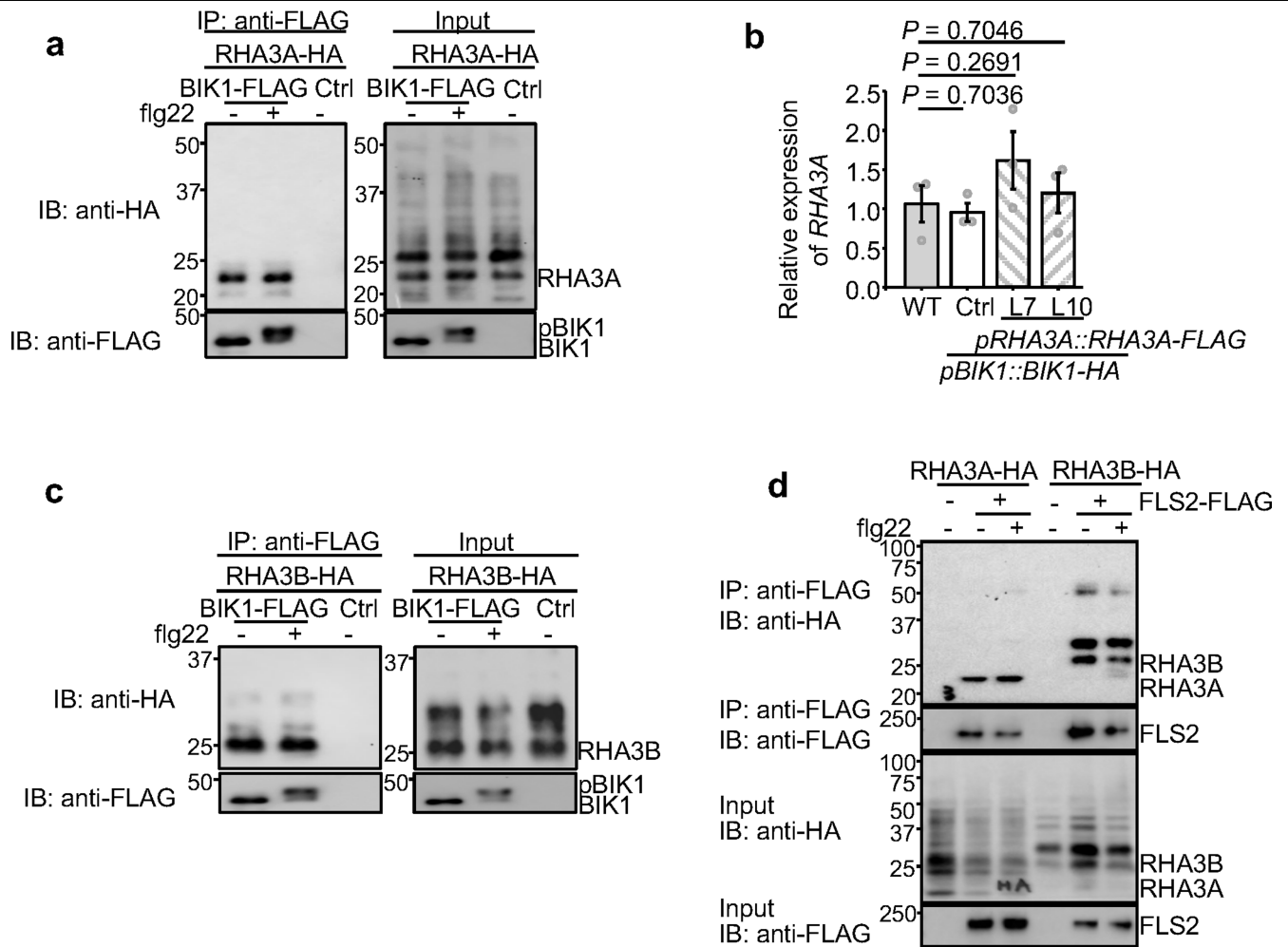
Extended Data Fig. 2 | MAMP-triggered monoubiquitination of BIK1-family RLCK proteins. **a**, Flg22 induces monoubiquitination of BIK1. Protoplasts from wild-type plants were transfected with BIK1-HA and FLAG-UBQ or a vector (Ctrl), and then treated with 100 nM flg22 for 30 min. After immunoprecipitation with anti-FLAG agarose, ubiquitinated BIK1 was detected by immunoblotting using anti-HA antibody (top). Middle, BIK1-HA proteins; bottom, CBB staining for RuBisCO (RBC). **b**, Flg22 induces BIK1 monoubiquitination in *pBIK1::BIK1-HA* transgenic plants. Protoplasts from *pBIK1::BIK1-HA/bik1* transgenic plants were transfected with FLAG-UBQ and then treated with 100 nM flg22 for 30 min. After immunoprecipitation with anti-FLAG agarose, ubiquitinated BIK1 was detected by immunoblotting with anti-HA antibody (top). Bottom, BIK1-HA proteins. **c**, BAK1 is constitutively polyubiquitinated in vivo. Protoplasts from wild-type plants were transfected with BAK1-HA and FLAG-UBQ or control, and then treated with 100 nM flg22 for 30 min. Immunoprecipitation was carried out with anti-FLAG agarose. Ub_n-BAK1 proteins were detected as a smear with anti-HA immunoblotting (top). Middle, BAK1-HA proteins; bottom, CBB staining for RBC. **d**, Flg22 induces FLS2 polyubiquitination. Protoplasts from wild-type plants were transfected with FLS2-HA and FLAG-UBQ and then treated with 100 nM flg22

for 30 min. **e**, **f**, Monoubiquitination of BIK1 with UBQ(K0). Protoplasts from *pBIK1::BIK1-HA* (**e**) or *35S::BIK1-HA* (**f**) transgenic plants were transfected with FLAG-UBQ(K0) (all lysine residues mutated to arginine) and then treated with 100 nM flg22 for 30 min. The mutations of lysine to arginine in UBQ(K0) are shown at the top of **e** with amino-acid positions labelled. **g**, PYR-41 blocks flg22-induced BIK1 monoubiquitination. PYR-41 (50 μM) was added 30 min before flg22 treatment. **h**, Flg22 induces BIK1 monoubiquitination in the presence of MG132. MG132 (2 μM) was added 1 h or 2.5 h before treatment with flg22. **i**, Flg22-induced BIK1 monoubiquitination depends on FLS2 and BAK1. Protoplasts isolated from wild-type, *fls2* or *bak1-4* plants were transfected with BIK1-HA and FLAG-UBQ and then treated with 100 nM flg22 for 30 min. **j**, elf18, pep1 and chitin induce BIK1 monoubiquitination. 1 μM elf18, 200 nM pep1 or 100 μg/ml chitin was added to protoplasts for 30 min. **k**, The BIK1 homologue PBL1 is monoubiquitinated upon treatment with flg22. PBL1-HA and FLAG-UBQ were expressed in protoplasts. **l**, Flg22 induces monoubiquitination of the BIK1-family RLCK PBL10 but not of BSK1. HA-tagged PBL10 or BSK1 was expressed with FLAG-UBQ in wild-type protoplasts. Experiments were repeated at least three times with similar results.



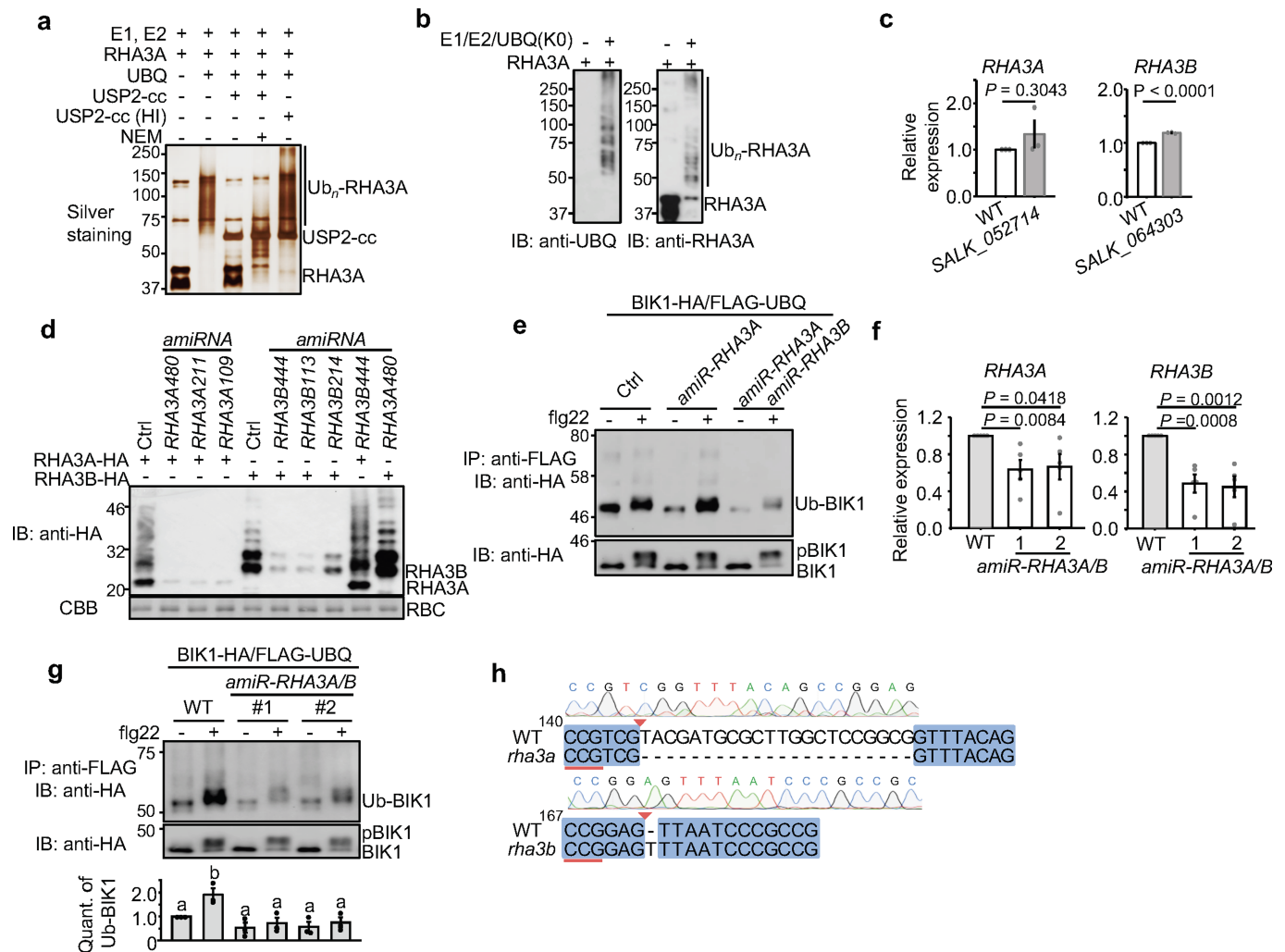
Extended Data Fig. 3 | Plasma membrane localization and phosphorylation are required for BIK1 ubiquitination. **a**, The kinase inhibitor K252a blocks flg22-induced ubiquitination of BIK1. Protoplasts transfected with FLAG-UBQ and BIK1-HA were treated with 1 μ M K252a for 30 min and then with 100 nM flg22. **b**, BIK1(G2A) no longer localizes to the plasma membrane. BIK1-YFP or BIK1(G2A)-YFP was expressed in *N. benthamiana* for imaging analysis. **c**, BIK1(G2A) show compromised flg22-induced monoubiquitination. BIK1-HA or BIK1(G2A)-HA was co-expressed with FLAG-UBQ in protoplasts. **d**, Single K-to-R mutations of BIK1 fail to block flg22-induced ubiquitination without

altering kinase activity. HA-tagged wild-type or mutant BIK1 was co-expressed with FLAG-UBQ in protoplasts. **e**, BIK1(K204R) exhibits reduced autophosphorylation and phosphorylation of BAK1. An in vitro kinase assay was performed using GST-BIK1 or GST-BIK1(K204R) as a kinase and GST or GST-BAK1^K (BAK1 kinase domain without detectable autophosphorylation activity) as a substrate with [γ -³²P]ATP. Top, proteins were separated with SDS-PAGE and analysed by autoradiography (Autorad.); bottom, protein loading shown CBB staining. Experiments were repeated at least twice with similar results.



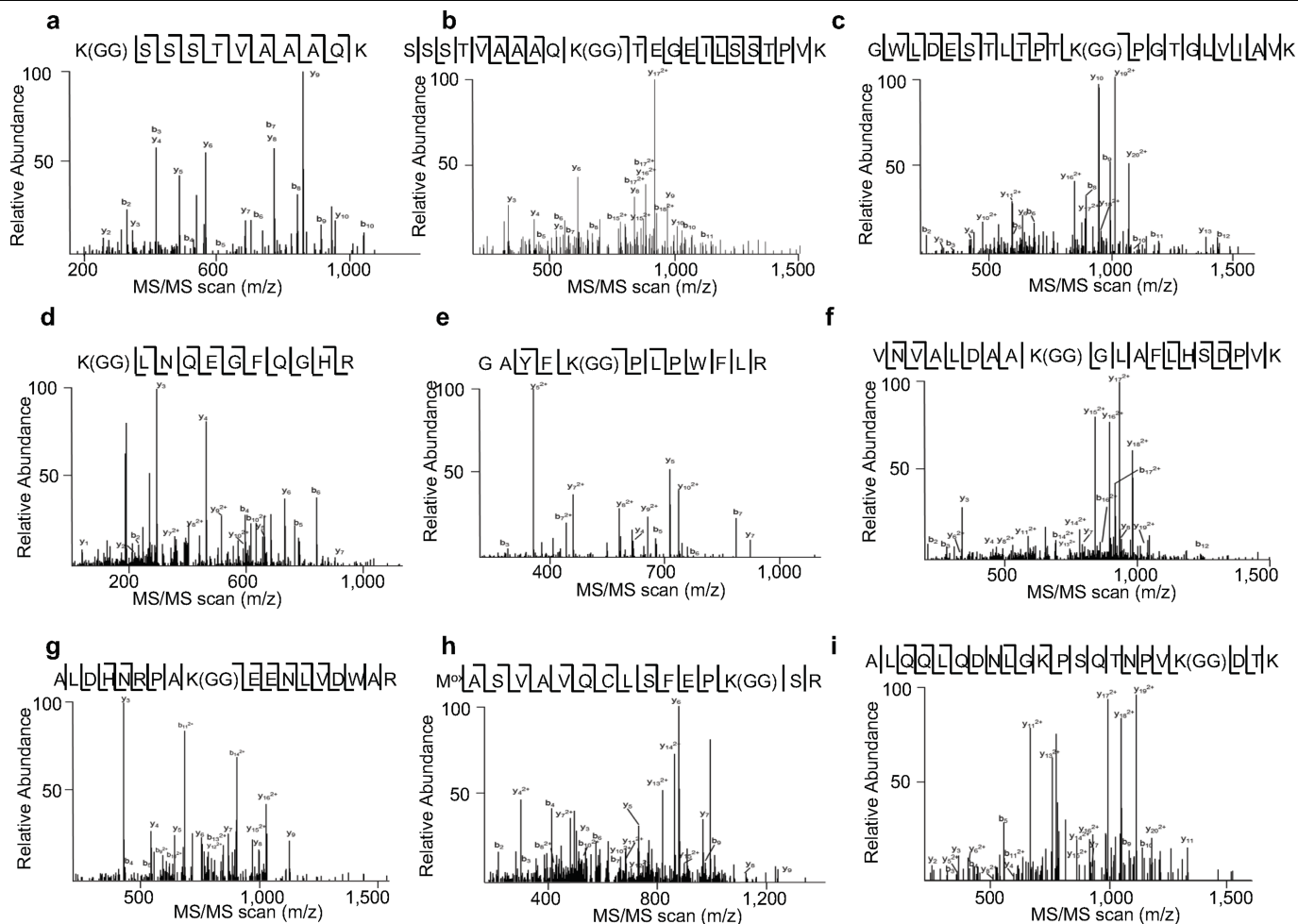
Extended Data Fig. 4 | RHA3A/B interacts with BIK1 in vivo. **a**, BIK1 interacts with RHA3A in a co-IP assay. RHA3A-HA was co-expressed with BIK1-FLAG or control in protoplasts and then treated with 100 nM flg22 for 15 min. Left, the co-IP assay was carried out with anti-FLAG agarose and immunoprecipitated proteins were immunoblotted with anti-HA or anti-FLAG antibody. Right, BIK1-FLAG and RHA3A-HA proteins. **b**, *RHA3A* expression (mean \pm s.e.m.) in *pRHA3A::RHA3A-FLAG/pBIK1::BIK1-HA* transgenic plants. qRT-PCR was carried out to detect *RHA3A* transcripts using *ACTIN2* as a control. Relative gene expression in wild-type (set as 1), *pBIK1::BIK1-HA* (Ctrl) and two independent

transgenic lines (lines 7 and 10) is shown. One-way ANOVA, $n = 3$. **c**, BIK1 associates with RHA3B independent of flg22 treatment. RHA3B-HA was co-expressed with BIK1-FLAG or control in protoplasts and then treated with 100 nM flg22 for 15 min. Left, co-IP assay was carried out with anti-FLAG agarose and immunoprecipitated proteins were immunoblotted with anti-HA or anti-FLAG antibody. Right, BIK1-FLAG and RHA3B-HA proteins before immunoprecipitation. **d**, FLS2 interacts with RHA3A and RHA3B in a co-IP assay. Experiments were repeated three times with similar results.



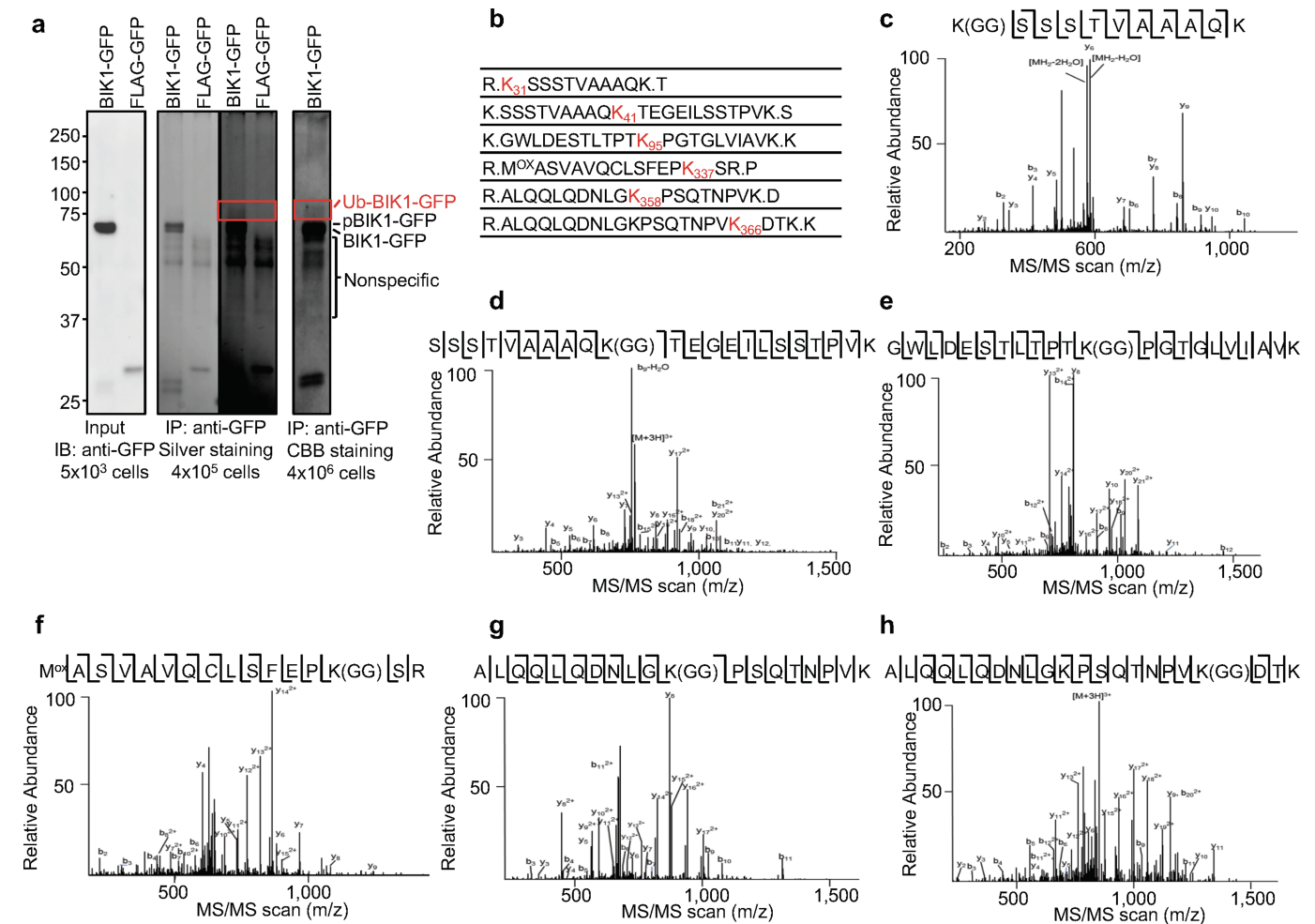
Extended Data Fig. 5 | RHA3A/B ubiquitinate BIK1 in vivo. **a**, GST-RHA3A^{CD} possesses E3 ligase activity in vitro. An in vitro ubiquitination assay was performed with GST-RHA3A^{CD} followed by deubiquitination reactions with GST-USP2-cc, *N*-ethylmaleimide (NEM) (10 mM), an inhibitor of deubiquitinases, and heat-inactivated (HI, 95 °C for 5 min) USP2-cc are controls. Samples were analysed by SDS-PAGE and silver staining. **b**, GST-RHA3A^{CD} possesses multi-monoubiquitination activity in vitro. A ubiquitination assay was done as in **a** but using the ubiquitin mutant with all lysine residues mutated to arginine (UBQ(K0)). Ubiquitinated proteins were detected by immunoblotting with anti-UBQ (left) or anti-RHA3A (right) antibodies. **c**, *RHA3* expression in T-DNA insertion mutants. *RHA3A* expression in the T-DNA knockout line *SALK_052714* and *RHA3B* expression in *SALK_064303* were analysed as in Extended Data Fig. 4b. Mean \pm s.e.m. fold change (WT set as 1.0); two-tailed Student's *t*-test, $n = 3$. **d**, Screen for the optimal *amiR-RHA3A* and *amiR-RHA3B*. Protoplasts were transfected with *RHA3A*-HA or *RHA3B*-HA with control, *amiR-RHA3A* or *amiR-RHA3B*. *RHA3A* or *RHA3B* proteins were examined by immunoblotting with anti-HA antibody. **e**, *RHA3A* and *RHA3B* are required for BIK1 ubiquitination in vivo. A BIK1 ubiquitination assay was carried out by co-expressing control, artificial microRNA targeting *RHA3A*

(*amiR-RHA3A*) or *amiR-RHA3A* together with microRNA targeting *RHA3B* (*amiR-RHA3A* *amiR-RHA3B*). **f**, *RHA3A* and *RHA3B* expression in *amiR-RHA3A/B* transgenic plants. qRT-PCR was carried out to detect *RHA3A* and *RHA3B* transcripts with *ACTIN2* as a control. Mean \pm s.e.m. fold change in gene expression from two independent transgenic lines (lines 1 and 2); one-way ANOVA, $n = 5$. *RHA3A* and *RHA3B* are required for BIK1 ubiquitination in transgenic plants. Protoplasts from *amiR-RHA3A/B* transgenic plants were transfected with BIK1-HA and FLAG-UBQ for ubiquitination assay. Bottom, quantification of BIK1 ubiquitination in *amiR-RHA3A/B* transgenic plants. Intensity of Ub-BIK1 or BIK1 bands was quantified with Image Lab (Bio-Rad). The amount of BIK1 ubiquitination is the relative intensity of the Ub-BIK1 band to the BIK1 band (no treatment in wild-type set as 1.0). Mean \pm s.e.m.; different letters indicate significant difference with others ($P < 0.05$, one-way ANOVA, $n = 3$). **h**, Sequencing analysis of *RHA3A* and *RHA3B* genes in the CRISPR-Cas9 *rha3a/b* mutant. PCR fragments corresponding to *RHA3A* and *RHA3B* in *rha3a/b* were amplified, sequenced, and aligned to wild-type coding sequences. The reverse complement of the PAM sequence is underlined in red, and red arrowheads indicate the theoretical Cas9 cleavage sites. The experiments were repeated three times with similar results.



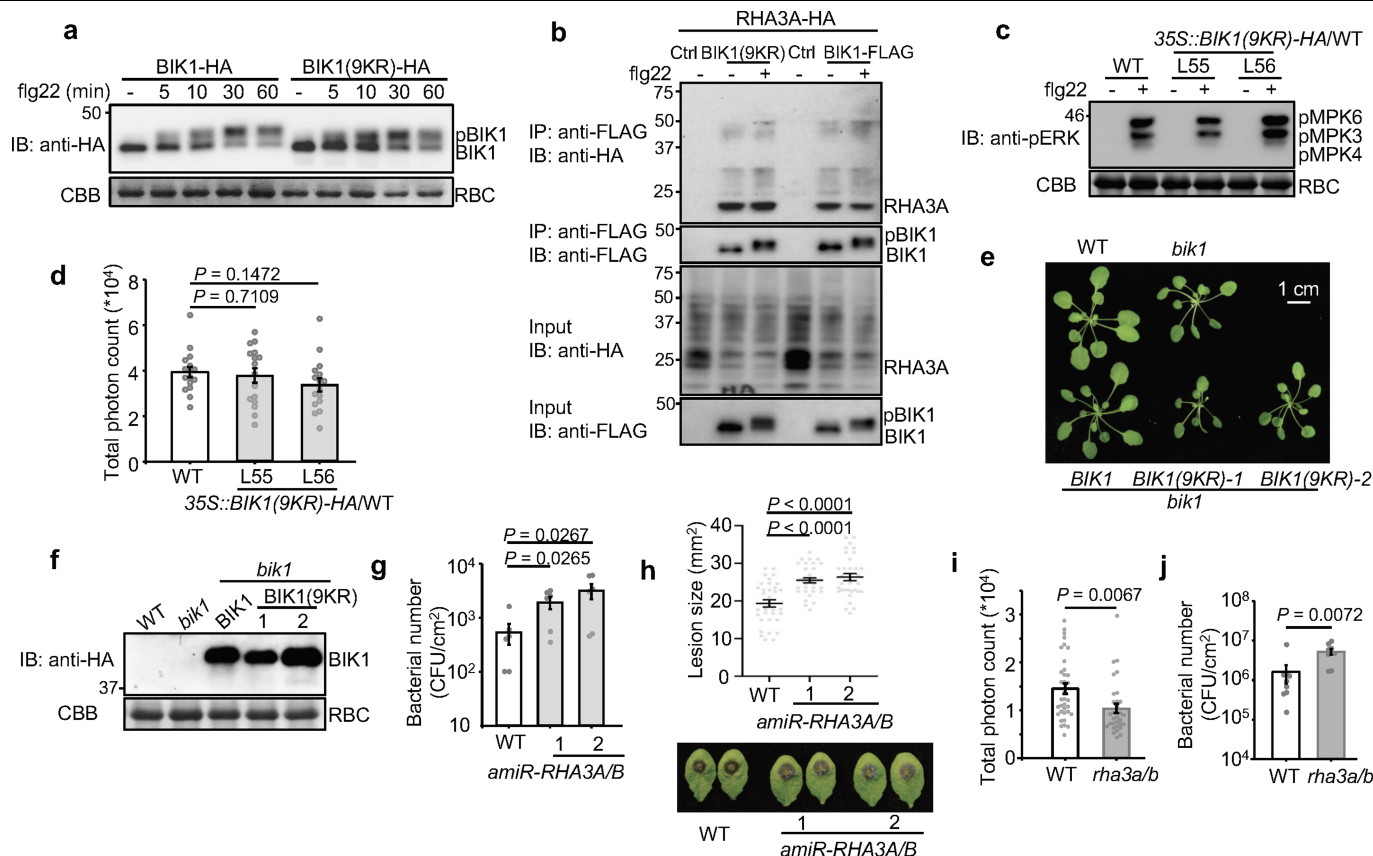
Extended Data Fig. 6 | BIK1 in vitro ubiquitination sites identified by mass spectrometry. MS/MS spectra of peptides containing ubiquitinated lysine residues of BIK1. **a**, K31; **b**, K41; **c**, K95; **d**, K106; **e**, K170; **f**, K186; **g**, K286; **h**, K337;

i, K366. MS spectra are outputs from the SEQUEST program. MS analysis was performed once.



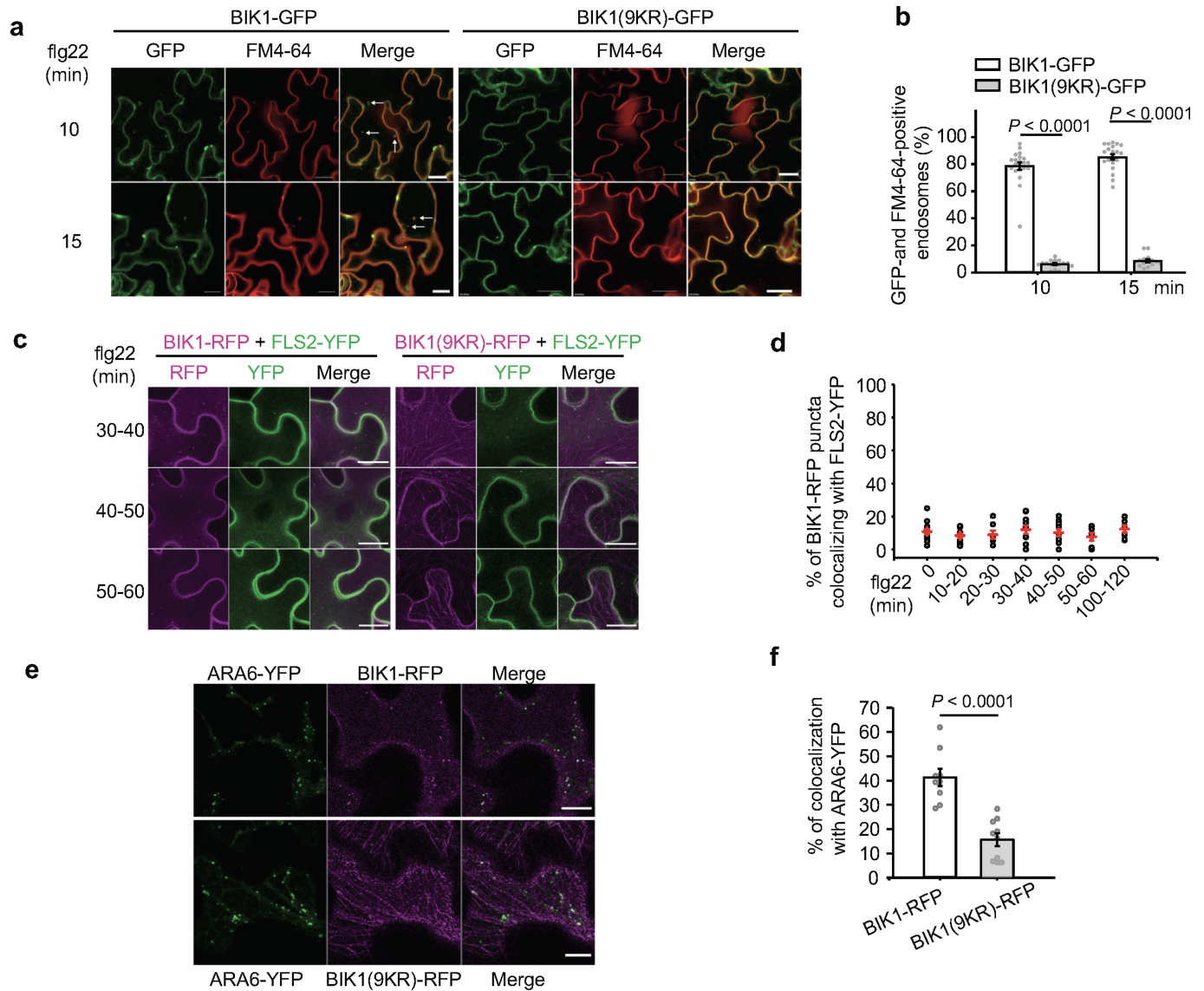
Extended Data Fig. 7 | BIK1 in vivo ubiquitination sites identified by mass spectrometry. a, Ubiquitinated BIK1-GFP in planta was immunoprecipitated for LC-MS/MS analysis. BIK1-GFP and FLAG-UBQ were co-expressed in wild-type protoplasts (about 4 × 10⁶ cells) and then treated with 200 nM flg22 for 30 min. Ubiquitinated BIK1 was immunoprecipitated with GFP-trap-agarose, separated by SDS-PAGE, digested with trypsin and subjected to LC-MS/MS analysis. Portions of cell lysates were examined for BIK1-GFP expression (left), and immunoprecipitates were analysed by SDS-

PAGE following silver staining (middle; right for longer exposure of the same gel) and SDS-PAGE following CBB staining (right). The highlighted area was cut and analysed by MS. **b**, BIK1 is ubiquitinated in vivo. Ubiquitinated lysines containing a diglycine remnant identified by LC-MS/MS analysis are marked in red with amino acid positions. **c-h**, MS/MS spectra of peptides containing ubiquitinated lysines of BIK1 are shown. **c**, K31; **d**, K41; **e**, K95; **f**, K337; **g**, K358; **h**, K366. MS spectra are outputs from the SEQUEST program. MS analysis was performed once.



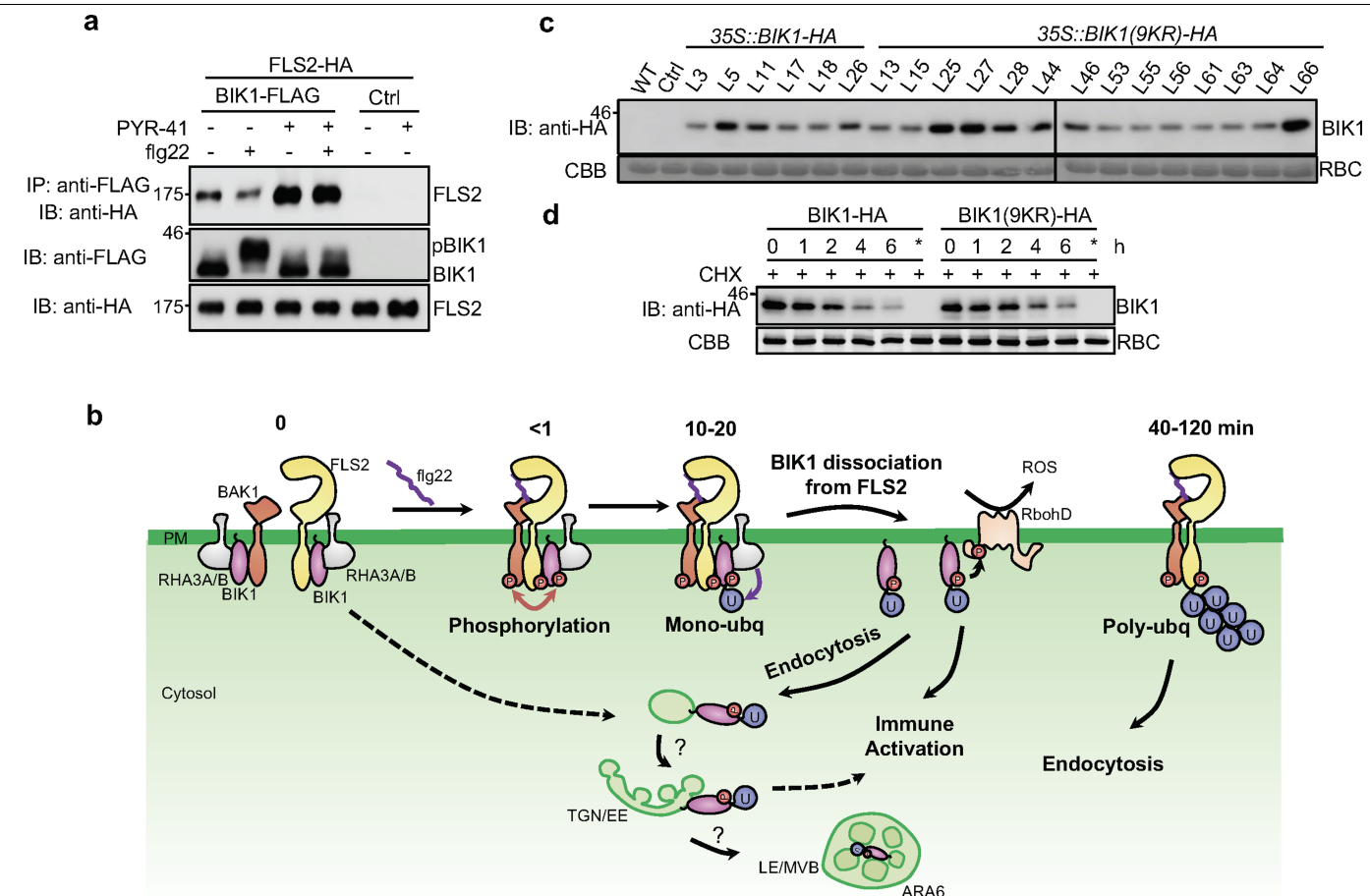
Extended Data Fig. 8 | BIK1 monoubiquitination is required for plant defence and flg22 signalling. **a**, BIK1(9KR) undergoes phosphorylation similar to BIK1 upon flg22 treatment. BIK1-HA or BIK1(9KR)-HA was expressed in wild-type protoplasts which were then treated with 100 nM flg22 for the indicated times. Band-shift of BIK1 was examined by immunoblotting with anti-HA antibody. **b**, BIK1(9KR) interacts with RHA3A in a co-IP assay. RHA3A-HA was co-expressed with BIK1-FLAG or BIK1(9KR)-FLAG in protoplasts that were then treated with 100 nM flg22 for 15 min. Co-IP assay was carried out with anti-FLAG agarose and immunoprecipitated proteins were immunoblotted with anti-HA or anti-FLAG antibody (top two panels). Bottom two panels show BIK1-FLAG or BIK1(9KR)-FLAG and RHA3A-HA proteins. **c**, Transgenic plants with *BIK1^{9KR}* overexpression in wild-type background show similar MAPK activation to wild-type plants. Eleven-day-old seedlings of wild-type or *35S::BIK1^{9KR}-HA/WT* transgenic plants (lines 55 and 56) were treated with 200 nM flg22 for 15 min. MAPK activation was analysed with anti-pERK antibody (top), and protein loading is shown by CBB staining for RBC (bottom). **d**, Transgenic plants with *BIK1^{9KR}* overexpression in wild-type background show similar flg22-induced ROS production to wild-type plants. Leaf discs from the indicated genotypes were treated with 100 nM flg22, and ROS production was measured as relative luminescence units by a luminometer over 50 min. Mean total photon count \pm s.e.m. overlaid with dot plot (one-way ANOVA, $n = 16$).

e, Growth phenotype of *pBIK1::BIK1-HA/bik1* and *pBIK1::BIK1^{9KR}-HA/bik1* transgenic plants. Five-week-old soil-grown plants are shown. Scale bar, 1 cm. **f**, Expression of BIK1-HA or BIK1(9KR)-HA in transgenic plants. Top, total proteins from leaves of four-week-old transgenic plants were subjected to anti-HA immunoblotting. Bottom, CBB staining for RBC. **g**, RHA3A and RHA3B are involved in resistance to *Pst* DC3000 *hrcC* infection. Plants were spray-inoculated with *Pst* DC3000 *hrcC* and bacterial growth was measured at 4 dpi. Mean \pm s.e.m. overlaid with dot plots (one-way ANOVA, $n = 6$). **h**, RHA3A and RHA3B are involved in resistance to *Botrytis*. Four-week-old plant leaves were deposited with 10 μ l *B. cinerea* BO5 at a concentration of 2.5×10^5 spores per ml. Disease symptoms were recorded, and the lesion diameter was measured at 2 dpi. Mean \pm s.e.m. overlaid with dot plots (one-way ANOVA, $n = 34$). **i**, ROS production is reduced in *rha3a/b* plants. Leaf discs from wild-type or *rha3a/b* plants were treated with 100 nM flg22 and ROS production measured over 50 min. Mean \pm s.e.m. total photon count overlaid with dot plots (two-tailed Student's *t*-test, $n = 36$ for wild-type and $n = 32$ for *rha3a/b*). **j**, RHA3A and RHA3B are involved in resistance to *Pst* DC3000. Plants were spray-inoculated with *Pst* DC3000 and bacterial growth was measured at 3 dpi. Mean \pm s.e.m. overlaid with dot plots (two-tailed Student's *t*-test, $n = 9$). Experiments were repeated three times with similar results.



Extended Data Fig. 9 | The BIK1(9KR) mutation impairs flg22-induced endocytosis of BIK1. **a, b**, BIK1(9KR)-GFP puncta colocalize less than BIK1-GFP with FM4-64 upon treatment with flg22. **a**, Five-day-old 35S::BIK1-GFP or 35S::BIK1^{9KR}-GFP seedlings were pretreated with FM4-64 (2 μ M) for 15 min and elicited with 100 nM flg22 for the indicated times; fluorescence was detected in epidermis using confocal microscopy. White arrows, colocalized endosomes. Scale bars, 20 μ m. **b**, Percentage of endosomes positive for BIK1-GFP or BIK1(9KR)-GFP and FM4-64 over time per 100% of image area. Mean \pm s.e.m. overlaid with dot plots (two-tailed Student's *t*-test, $n = 21$ images for BIK1-GFP and $n = 16, 15$ images for 10, 15 min, respectively, for BIK1(9KR)-GFP). **c**, Flg22-induced endocytosis of BIK1, BIK1(9KR) and FLS2 in *N. benthamiana*. BIK1-TagRFP (BIK1-RFP) or BIK1(9KR)-TagRFP (BIK1(9KR)-RFP) was co-expressed with FLS2-YFP in *N. benthamiana*, infiltrated with 100 μ M flg22 and imaged at the indicated time points by confocal microscopy. Images at 30-40, 40-50 and 50-60 min after flg22 treatment from Fig. 4e are shown

here. Scale bars, 20 μ m. For BIK1-RFP/FLS2-YFP, $n = 14, 11, 7, 10, 10, 6, 7$ images for 0, 10-20, 20-30, 30-40, 40-50, 50-60, 100-120 min; for BIK1(9KR)-RFP/FLS2-YFP, $n = 19, 11, 11, 9, 16, 12, 7$ images for 0, 10-20, 20-30, 30-40, 40-50, 50-60, 100-120 min, respectively. **d**, Percentage of BIK1-RFP puncta that colocalized with FLS2-YFP after treatment with flg22 for the indicated times in **c** and Fig. 4e. Mean \pm s.e.m. overlaid with dot plots ($n = 14, 11, 7, 10, 10, 6, 7$ images for 0, 10-20, 20-30, 30-40, 40-50, 50-60, 100-120 min, respectively). **e, f**, BIK1(9KR)-RFP shows reduced colocalization with ARA6-YFP. **e**, BIK1-RFP or BIK1(9KR)-RFP was transiently expressed with ARA6-YFP in *N. benthamiana*, and the images were taken 48-72 h after infiltration. Scale bars, 10 μ m. **f**, Percentage of BIK1-RFP puncta that colocalized with ARA6-YFP. Mean \pm s.e.m. overlaid with dot plots (two-tailed Student's *t*-test, $n = 9$ images for BIK1-RFP; $n = 10$ images for BIK1(9KR)-RFP). Experiments were repeated three times with similar results.



Extended Data Fig. 10 | Monoubiquitination mediates release of BIK1 from the plasma membrane upon ligand detection. a, PYR-41 impairs flg22-induced dissociation of BIK1 from FLS2. FLS2-HA was co-expressed with BIK1-FLAG or control in protoplasts. After pretreatment with 50 μ M PYR-41 for 30 min, protoplasts were stimulated with 100 nM flg22 for 15 min. Co-IP and immunoblotting were performed as in Fig. 4g. **b**, A working model of RHA3A/B-mediated BIK1 monoubiquitination in plant immunity. Under non-activated, steady-state conditions (0 min), BIK1 remains hypo-phosphorylated and associates with FLS2 and BAK1. Upon flg22 detection, FLS2 dimerizes with BAK1, which stimulates BIK1 phosphorylation (<1 min). Phosphorylated BIK1 is monoubiquitinated by the E3 ligases RHA3A and RHA3B, leading to dissociation of BIK1 from the FLS2-BAK1 complex, accompanied by endocytosis (10–20 min). Ligand-induced

monoubiquitination of BIK1 contributes to the activation of ROS and other defence responses. FLS2 is polyubiquitinated and endocytosed 40 min after detection of flg22 to attenuate signalling. **c**, BIK1(9KR) shows comparable protein expression to BIK1 in transgenic plants. 35S::BIK1-HA or 35S::BIK1^{9KR}-HA transgenic plants in wild-type background were used for immunoblotting to detect BIK1 proteins with anti-HA antibody. Control, empty vector. **d**, Stability of BIK1 and BIK1(9KR) proteins after treatment with cycloheximide (CHX). BIK1-HA or BIK1(9KR)-HA was expressed in wild-type protoplasts for 12 h followed by treatment with 500 μ g/ml CHX for the indicated time. BIK1 or BIK1(9KR) proteins were analysed by immunoblotting with anti-HA antibody. Asterisk indicates that CHX was added immediately after transfection, thus blocking protein synthesis. Experiments were repeated three times with similar results.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Image J (version 1.51), Image Lab (Bio-Rad, version 4.1), LAS-X (Leica, version 3.5.6.21594) were used in data collection.

Data analysis

Images from Immuno Blotting were quantified with Image LabTM (Bio-Rad, version 4.1). Confocal images were analyzed with LAS-X (Leica, version 3.5.6.21594), ZEN (Zeiss). Statistical analysis was performed with Microsoft Excel 2016. The MS/MS spectra were analyzed with SEQUEST (version 28). Crystal structure was analyzed with PyMOL (version 2.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data supporting the findings of this study are available within the paper and its Supplementary Information files. Source Data (gels and graphs) for Figs. 1–4 and Extended Data Figs. 1–10 are provided with the paper.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes. Sample sizes were determined based on previous publications on similar experiments. The sample sizes were sufficient as the differences between experimental groups were reproducible.
Data exclusions	No data were excluded from analyses in the experiments.
Replication	All attempts to replicate the experiments were successful. Number of repeats was given in the figure legends.
Randomization	Plant materials used in the study were collected randomly.
Blinding	Investigators were not blinded to plant genotypes during experiments. The research materials are plants so the blinding design is not applicable to this system. Experiment results are not subjective.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.
Did the study involve field work?	<input type="checkbox"/> Yes <input type="checkbox"/> No

Field work, collection and transport

Field conditions	Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).
Location	State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).
Access and import/export	Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).
Disturbance	Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials

Describe any restrictions on the availability of unique materials OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources).

Antibodies

Antibodies used

Anti-HA-Peroxidase, Roche, Cat # 12013819001, clone 3F10. Dilution 1: 2,000.
 Anti-FLAG-Peroxidase, Sigma-Aldrich, Cat # A8592, clone M2. Dilution 1: 2,000.
 Anti-GST, Santa Cruz, Cat # SC-53909, clone 1E5. Dilution 1: 2,000.
 Anti-ubiquitin (P4D1), Santa Cruz, Cat # SC-8017, clone P4D1. Dilution 1:500.
 Anti-GFP, Roche, Cat # 11814460001, mix of clone 7.1 and 13.1. Dilution 1: 2,000.
 Anti-Phospho-p44/42 MAPK, Cell Signaling, Cat # 9101. Polyclone. Dilution 1: 2,000.
 Anti-Rabbit IgG, HRP-linked antibody, Cell Signaling, Cat # 7074. Polyclone. Dilution 1: 10,000.
 Anti-Mouse IgG, HRP-linked antibody, Cell Signaling, Cat # 7076. Polyclone. Dilution 1: 10,000.
 Anti-FLAG M2 Affinity gel, Sigma-Aldrich, Cat # 2220, clone M2.
 Anti-MBP, Biolegend, Cat # 906901, clone YM-2. Dilution 1: 1,000.
 Anti-RHA3A, GenScript, generated with peptide AGGDSPSPNKGKLC. Polyclone. Dilution 1:1,000.

Validation

Validation statements, relevant citations of commercial primary antibodies are available from manufacturers:
 Anti-HA-Peroxidase, Roche, Cat # 12013819001, clone 3F10. <https://www.sigmaaldrich.com/catalog/product/roche/12013819001?lang=en®ion=US>
 Anti-FLAG-Peroxidase, Sigma-Aldrich, Cat # A8592, clone M2. <https://www.sigmaaldrich.com/catalog/product/sigma/a8592?lang=en®ion=US>
 Anti-GST, Santa Cruz, Cat # SC-53909, clone 1E5. <https://www.scbt.com/p/gst-antibody-1e5>
 Anti-ubiquitin (P4D1), Santa Cruz, Cat # SC-8017, clone P4D1. <https://www.scbt.com/p/ub-antibody-p4d1>
 Anti-GFP, Roche, Cat # 11814460001, mix of clone 7.1 and 13.1. <https://www.sigmaaldrich.com/catalog/product/roche/11814460001?lang=en®ion=US>
 Anti-Phospho-p44/42 MAPK, Cell Signaling, Cat # 9101. <https://www.cellsignal.com/products/primary-antibodies/phospho-p44-42-mapk-erk1-2-thr202-tyr204-antibody/9101>
 Anti-Rabbit IgG, HRP-linked antibody, Cell Signaling, Cat # 7074. <https://www.cellsignal.com/products/secondary-antibodies/anti-rabbit-igg-hrp-linked-antibody/7074>
 Anti-Mouse IgG, HRP-linked antibody, Cell Signaling, Cat # 7076. <https://www.cellsignal.com/products/secondary-antibodies/anti-mouse-igg-hrp-linked-antibody/7076>
 Anti-FLAG M2 Affinity gel, Sigma-Aldrich, Cat # 2220, clone M2. <https://www.sigmaaldrich.com/catalog/product/sigma/a2220?lang=en®ion=US>
 Anti-MBP, Biolegend, Cat # 906901, clone YM-2. <https://www.biolegend.com/en-gb/products/purified-anti-maltose-binding-protein-mbp-antibody-11081>
 Anti-RHA3A, GenScript, generated with peptide AGGDSPSPNKGKLC from Rabbit. Figure 2d bottom panel supports the anti-RHA3A antibody do not have cross reactivity with other protein including E1, E2, BIK1 or Ubiquitin.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

State the source of each cell line used.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See [ICLAC](#) register)

No commonly misidentified cell lines were used.

Palaeontology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

- Sample preparation *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
- Instrument *Identify the instrument used for data collection, specifying make and model number.*
- Software *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
- Cell population abundance *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
- Gating strategy *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*
- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

- Design type *Indicate task or resting state; event-related or block design.*
- Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
- Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

- Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*
- Field strength *Specify in Tesla*
- Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
- Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*
- Diffusion MRI ☐ Used ☐ Not used

Preprocessing

- Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
- Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
- Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
- Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
- Volume censoring *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See [Eklund et al. 2016](#))

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a | Involved in the study

- ☐ ☐ Functional and/or effective connectivity
☐ ☐ Graph analysis
☐ ☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.

Brain control of humoral immune responses amenable to behavioural modulation

<https://doi.org/10.1038/s41586-020-2235-7>

Received: 7 December 2018

Accepted: 20 March 2020

Published online: 29 April 2020

 Check for updates

Xu Zhang^{1,2,3,4,16}, Bo Lei^{4,5,16}, Yuan Yuan^{6,16}, Li Zhang^{1,2,3,4,16}, Lu Hu⁷, Sen Jin⁸, Bilin Kang^{4,5}, Xuebin Liao⁷, Wenzhi Sun^{9,10}, Fuqiang Xu^{8,11,12}, Yi Zhong^{4,5}, Ji Hu^{6,13} & Hai Qi^{1,2,3,4,14,15}✉

It has been speculated that brain activities might directly control adaptive immune responses in lymphoid organs, although there is little evidence for this. Here we show that splenic denervation in mice specifically compromises the formation of plasma cells during a T cell-dependent but not T cell-independent immune response. Splenic nerve activity enhances plasma cell production in a manner that requires B-cell responsiveness to acetylcholine mediated by the $\alpha 9$ nicotinic receptor, and T cells that express choline acetyl transferase^{1,2} probably act as a relay between the noradrenergic nerve and acetylcholine-responding B cells. We show that neurons in the central nucleus of the amygdala (CeA) and the paraventricular nucleus (PVN) that express corticotropin-releasing hormone (CRH) are connected to the splenic nerve; ablation or pharmacogenetic inhibition of these neurons reduces plasma cell formation, whereas pharmacogenetic activation of these neurons increases plasma cell abundance after immunization. In a newly developed behaviour regimen, mice are made to stand on an elevated platform, leading to activation of CeA and PVN CRH neurons and increased plasma cell formation. In immunized mice, the elevated platform regimen induces an increase in antigen-specific IgG antibodies in a manner that depends on CRH neurons in the CeA and PVN, an intact splenic nerve, and B cell expression of the $\alpha 9$ acetylcholine receptor. By identifying a specific brain–spleen neural connection that autonomically enhances humoral responses and demonstrating immune stimulation by a bodily behaviour, our study reveals brain control of adaptive immunity and suggests the possibility to enhance immunocompetency by behavioural intervention.

We developed a surgical denervation protocol by treating splenic nerve plexuses with alcohol before they enter the spleen along the vasculature (Extended Data Fig. 1a–c). The tyrosine hydroxylase-containing fibres were normally seen in the T cell zone, the T–B border and bridging channels, but disappeared after denervation (Fig. 1a, Extended Data Fig. 1d–g). Mice with denervated spleens were grossly normal and survived for as long as did sham-operated mice. Norepinephrine levels in spleen homogenates were reduced after denervation (Extended Data Fig. 1h). No abnormal lymphocyte apoptosis was seen in denervated spleens (Extended Data Fig. 1i). Six weeks after surgery, we immunized denervated and sham-operated mice intraperitoneally with 100 μ g NP-KLH (4-hydroxy-3-nitrophenylacetyl hapten conjugated to keyhole limpet haemocyanin) in alum plus 1 μ g lipopolysaccharide (LPS) and characterized germinal centre (GC) development and splenic plasma cell

(SPPC) formation 7 or 13 days later. The gating strategy (singlets→live events→total B-lineage cells) is shown in Fig. 1b. To quantify frequencies of GCs and SPPCs, we used total B-lineage cells including both CD19⁺ cells and CD138⁺ plasma cells for normalization. We did not detect significant differences in GC formation between the two groups (Fig. 1c). From day 7 to 13, the SPPC frequency in total B-lineage cells increased in sham-operated but not denervated mice (Fig. 1d). Notably, when the same mice were immunized with the T-independent antigen NP-Ficoll, we found no difference in SPPC formation (Extended Data Fig. 1j, k). Given the comparably normal tissue anatomy, the GC response, and T-independent SPPC formation in denervated mice, it is not likely that defects in the T-dependent response are caused by disrupted organ physiology. Instead, splenic nerve activity might specifically regulate T-dependent SPPC formation.

¹Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing, China. ²Laboratory of Dynamic Immunobiology, Institute for Immunology, Tsinghua University, Beijing, China.

³Department of Basic Medical Sciences, School of Medicine, Tsinghua University, Beijing, China. ⁴School of Life Sciences, Tsinghua University, Beijing, China. ⁵McGovern Institute of Brain Research, Beijing, China. ⁶School of Life Science and Technology, ShanghaiTech University, Shanghai, China. ⁷School of Pharmacological Sciences, Tsinghua University, Beijing, China.

⁸Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ⁹School of Basic Medical Sciences, Capital Medical University, Beijing, China. ¹⁰Chinese Institute for Brain Research, Beijing, China. ¹¹Centre for Brain Science, State Key Laboratory of Magnetic Resonance and Atomic Molecular Physics, Key Laboratory of Magnetic Resonance in Biological Systems, Wuhan Institute of Physics and Mathematics, Wuhan, China. ¹²Centre for Excellence in Brain Science and Intelligent Technology, Chinese Academy of Sciences, Wuhan, China. ¹³Co-Innovation Center of Neuroregeneration, Nantong University, Nantong, China. ¹⁴Beijing Key Laboratory for Immunological Research on Chronic Diseases, Tsinghua University, Beijing, China. ¹⁵Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, China. ¹⁶These authors contributed equally: Xu Zhang, Bo Lei, Yuan Yuan, Li Zhang.

✉e-mail: zhongyi@tsinghua.edu.cn; huji@shanghaitech.edu.cn; qihai@tsinghua.edu.cn

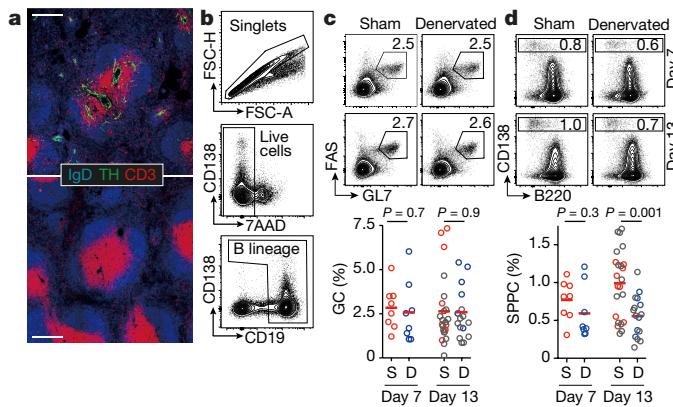


Fig. 1 | Splenic denervation reduces the abundance of plasma cells after immunization with T-dependent antigen. **a**, Images of sham-operated (top) or denervated (bottom) spleens 6 weeks after surgery, representative of three experiments. Green, TH staining (nerve fibres); blue, IgD staining (follicle); red, CD3 ϵ (T cell zone). Scale bar, 200 μ m. **b**, Gating scheme for quantifying GC and SPPC frequencies. **c**, **d**, Representative contour plots (top) and summary data (bottom) of frequencies of Fas^{hi}GL7^{hi} GC B cells (**c**) and CD138⁺ SPPCs (**d**) 7 or 13 days after immunization. Summary data pooled from five independent experiments; each symbol indicates one mouse, lines indicate means. In two experiments both day 7 and day 13 were examined (red, blue), and in the other three only day 13 was examined (grey). Two-tailed unpaired *t*-test. S, sham-operated; D, denervated.

The splenic nerve releases norepinephrine. In response to adrenergic stimulation, non-neural elements such as choline acetyl transferase (ChAT)-expressing T cells can synthesize and release acetylcholine in the spleen^{1,2}. By intravenous infusion of norepinephrine or acetylcholine during immunization, and by transduction of B cells with membrane-anchored acetylcholinesterase, we found that developing SPPCs were likely to be exposed to acetylcholine and subjected to its direct influence *in vivo* (Supplementary Text 1 and Extended Data Fig. 2).

We therefore further tested the expression of acetylcholine receptors by B cells. There are two classes of acetylcholine receptor (AChR): G-protein-coupled muscarinic receptors and pentameric, nicotinic ligand-gated channels. B-lineage cells did not express substantial amounts of muscarinic AChRs or nicotinic AChRs containing the $\alpha 7$ subunit, which is highly expressed by myeloid cells^{3,4}; however, they abundantly expressed nicotinic AChRs containing the $\alpha 9$, $\beta 1$, and $\beta 4$ subunits (Fig. 2a). Next, we used the CRISPR–Cas9 gene editing technique to generate mice lacking the *Chrna9*, *Chrn1* and *Chrn4* genes on the B6 background. No homozygous germline *Chrn1* knockout was obtained, potentially indicating embryonic lethality (data not shown). However, both *Chrna9*^{−/−} and *Chrn4*^{−/−} mice were viable and backcrossed with wild-type B6 mice for three generations (Extended Data Fig. 3a–d). Upon interbreeding, *Chrn4*^{−/−} mice had only infrequent litters, which contained fewer pups than normal (data not shown). *Chrna9*^{−/−} mice were fertile, born with the expected Mendelian frequency, and contained normal B cell and T cell compartments (Extended Data Fig. 3e, f and data not shown). Using a synthetic analogue of acetylcholine (Extended Data Fig. 3g), we measured the acetylcholine-binding capacity of B-lineage cells from wild-type and *Chrna9*^{−/−} mice. SPPCs exhibited a higher capacity to bind acetylcholine than GC or total B cells, and an $\alpha 9$ deficiency led to a substantial reduction in the acetylcholine-binding capacity of SPPCs (Extended Data Fig. 3h, i).

Bone marrow cells from *Chrna9*^{−/−} and μ MT (B cell-deficient) mice were mixed at a 20:80 ratio and transplanted into radiation-treated wild-type mice to create chimaeras in which all B cells were deficient in the $\alpha 9$ nicotinic AChR subunit, whereas other haematopoietic cells were largely intact. Two weeks after immunization of these chimaeric mice with NP-KLH, *Chrna9*^{−/−} B cells exhibited a significant defect in

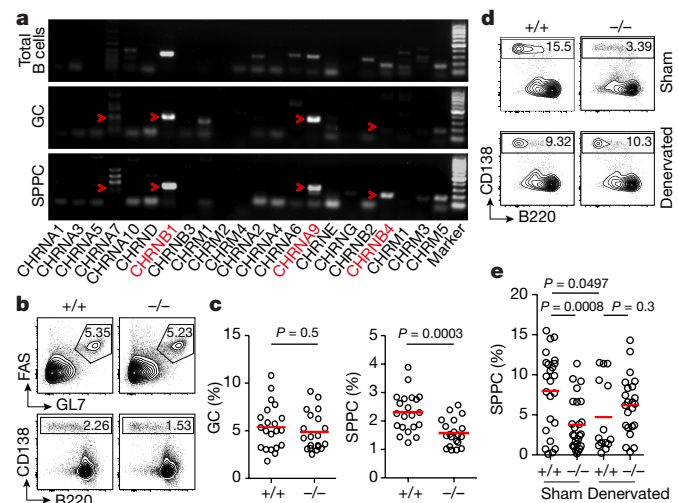


Fig. 2 | B cell-intrinsic responsiveness to acetylcholine underlies plasma cell formation promoted by splenic nerve activities. **a**, Expression of nicotinic acetylcholine receptor subunits in sorted total B cells (7AAD[−]CD19⁺), GC B cells (7AAD[−]CD19⁺GL7⁺Fas⁺) and SPPCs (7AAD[−]CD138⁺). Arrowheads, correct $\alpha 7$, $\beta 1$, $\alpha 9$, and $\beta 4$ amplicon sizes. **b**, **c**, Representative contour plots (**b**) and summary data (**c**) of percentage SPPC and GC in bone-marrow chimaeras reconstituted with 80% μ MT and 20% *Chrna9*^{+/+} (+/+) or *Chrna9*^{−/−} (−/−) bone-marrow cells, 13 days after immunization. Data pooled from four independent experiments. Each symbol indicates one mouse; lines denote means. **d**, **e**, Representative contour plots (**d**) and summary data (**e**) of percentage SPPC in sham-operated or denervated μ MT mice that were reconstituted with intravenously infused *Chrna9*^{+/+} or *Chrna9*^{−/−} mature B cells. Each symbol indicates one mouse; lines denote means. Two-tailed unpaired *t*-tests (**c**, **e**).

generating SPPCs as compared to wild-type B cells, whereas the magnitude of the GC response was not altered (Fig. 2b, c). In addition, we tested μ MT mice directly infused with *Chrna9*^{−/−} B cells in three experiments and with *Chrn4*^{−/−} B cells in two experiments; in this setup, B cells deficient in either $\alpha 9$ or $\beta 4$ AChRs produced fewer SPPCs after immunization with NP-KLH than did wild-type control cells (Extended Data Fig. 4a, b). Therefore, acetylcholine promotes SPPC formation via nicotinic AChRs that contain the $\alpha 9$ subunit and probably the $\beta 4$ subunit. To investigate whether this acetylcholine effect depends on splenic nerve activity, we conducted splenic denervation on μ MT mice before reconstitution with wild-type or *Chrna9*^{−/−} B cells (Extended Data Fig. 4c). Whereas the loss of $\alpha 9$ AChRs on B cells led to a reduction in SPPCs in sham-operated mice, this reduction was abrogated in mice without intact splenic nerves (Fig. 2d, e). Together, these data indicate that splenic nerve activity promotes SPPC formation through an acetylcholine-induced process in B cells.

Activated CD4 T cells can express ChAT and secrete acetylcholine in response to norepinephrine¹, and ChAT-expressing T cells may translate noradrenergic signalling by the splenic nerve into acetylcholine-dependent promotion of SPPC formation. We reconstituted T cell-deficient mice with CD4 T cells isolated from ChAT-IRES-Cre:Rosa26^{Ai14} reporter mice (Extended Data Fig. 5a). No Ai14⁺ T cells were detected in the spleen before immunization with NP-KLH, but 8–10 days after NP-KLH immunization the frequency of Ai14⁺ CD4 T cells was about 0.5%. These Ai14⁺ cells were of a CD44^{hi} activated phenotype and expressed *Chat* mRNA (Extended Data Fig. 5b). On tissue sections, many of these cells were in regions rich in tyrosine hydroxylase (TH)-positive nerve fibres and aggregates of plasma cells (Extended Data Fig. 5c). Next, we used CD4 T cells from ChAT-IRES-Cre:Rosa26^{DTR} or control Rosa26^{DTR} mice to reconstitute T cell-deficient mice. One week after immunization with NP-KLH, these mice were given intraperitoneal injections of 50 μ g kg^{−1} diphtheria

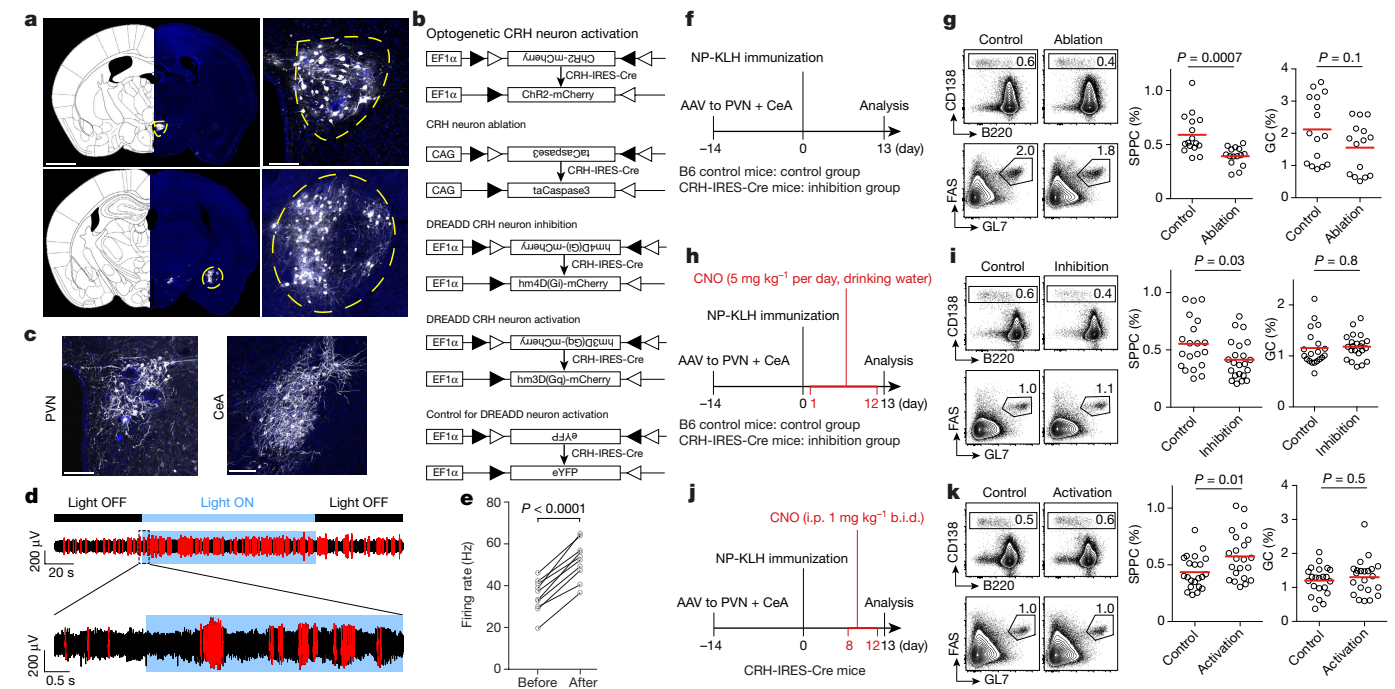


Fig. 3 | CRH neurons in CeA and PVN promote SPPC formation.

a, Representative coronal brain sections showing mRFP-labelled PVN (top) and CeA (bottom) from mice injected with PRV-mRFP in the spleen 96 h previously. Left, whole-brain views overlaid on the mouse brain atlas (scale bar, 1 mm); right, magnified views of the circled PVN and CeA regions (scale bar, 100 μ m). White, PRV-mRFP; blue, DAPI. **b**, AAV constructs used to manipulate CRH neurons. DIO-ChR2 for optogenetic activation, DIO-Caspase 3 for cell ablation, DIO-hM4D(Gi) for pharmacogenetic inhibition, DIO-hM3D(Gq) for pharmacogenetic activation, and DIO-fluorescent protein as infection control. **c**, Representative images of PVN and CeA regions from a CRH-IRES-Cre mouse injected with AAV-DIO-eYFP; white, YFP; blue, DAPI. Scale bars, 100 μ m. **d**, A representative trace of electrical activity in the splenic nerve before and during ChR2-mediated activation of CeA or PVN CRH neurons. Spikes >120 μ V are marked in red. **e**, Firing rates of the splenic nerve before and after optogenetic

stimulation of CeA and PVN CRH neurons ($n = 10$ trials, 5 male mice, 3 experiments; two-tailed paired t -test). **f**, **g**, Immunization following ablation of PVN and CeA CRH neuron ablation. **f**, Protocol (NP-KLH immunization is intraperitoneal (i.p.)); **g**, representative contour plots and summary data of SPPC and GC formation. Each symbol denotes one mouse; lines denote means; pooled data from three experiments. **h**, **i**, Immunization while PVN and CeA CRH neurons are inhibited. **h**, Protocol; **i**, representative contour plots and summary data of SPPC and GC formation. Each symbol denotes one mouse; lines denote means; pooled data from four experiments. **j**, **k**, Immunization while PVN and CeA CRH neuron activity is enhanced. **j**, Protocol; **k**, representative contour plots and summary data of SPPC and GC formation. Each symbol denotes one mouse; lines denote means; pooled data from four experiments. Two-tailed unpaired t -test (**g**, **i**, **k**).

toxin daily for 6 days to ablate DTR-expressing CD4 T cells. As shown in Extended Data Fig. 5d, e, significantly fewer SPPCs were generated in the ChAT-IRES-Cre;Rosa26^{DTR} group. GC responses were not significantly altered, indicating that the deletion procedure did not cripple the helper T cell compartment. These data support the possibility that ChAT-expressing T cells serve as a relay between the sympathetic splenic nerve and the acetylcholine-responsive process of SPPC formation. Other cell types may also contribute acetylcholine in this regard, although we did not find evidence to implicate B-cell-derived acetylcholine (Extended Data Fig. 5f–h).

To trace the brain origin of splenic nerve-dependent immunostimulatory neural activity, we conducted retrograde tracing by injecting fluorescent protein-expressing recombinant pseudo-rabies virus (PRV) Bartha strain into the spleen⁵. Over 96 h, PRV ascended from the spleen up into the spinal cord, brain stem and hypothalamus (Supplementary Tables 1, 2, Extended Data Fig. 6a–d), similar to findings in rats⁶. In the forebrain, the CeA and PVN were prominently labelled (Fig. 3a). The CeA orchestrates physiological and behavioural responses to threats and fears⁷, while the PVN processes inputs about stress and controls the hypothalamo-pituitary-adrenocortical (HPA) response⁸. Both CeA and PVN harbour abundant neurons that produce CRH⁹, the upstream hormone that drives the HPA axis to produce glucocorticoids^{10,11}. We hypothesized that, in contrast to the blood-borne effects of the HPA axis, the activity of CRH neurons might also be transmitted as efferent outputs through the splenic nerve to promote SPPC formation.

To test for a connection between CRH neurons in the CeA and PVN and the splenic nerve, we stereotactically injected an adeno-associated virus (AAV) that conditionally expresses Channelrhodopsin-2 (ChR2) and mCherry into the CeA and PVN regions of CRH-IRES-Cre knock-in mice (Fig. 3b, optogenetic setup). When triggered by light, ChR2 activates neurons in a Cre-dependent manner. After verification of correct stereotactic targeting (Fig. 3c), we conducted direct electrophysiological recordings from the splenic nerve during optogenetic activation of CRH neurons in the CeA or PVN (Extended Data Fig. 6e–g). Light stimulation of CRH neurons induced markedly increased firing through the splenic nerve (Fig. 3d, e). Therefore, CRH neurons in the CeA and PVN are connected to and can stimulate the splenic nerve.

To test the functional effects of CRH neuron activity on SPPC formation, we stereotactically injected an AAV that conditionally expresses death-inducing active caspase 3 into the CeA and PVN of CRH-IRES-Cre mice (Fig. 3b, ablation setup). As assessed with the Rosa26^{Al3} reporter line, our targeting and ablation efficiency of CRH neurons was about 80% in both the CeA and the PVN (Extended Data Fig. 7a–c). Following immunization, SPPC formation was significantly impaired in the CRH ablation group as compared to the control group (Fig. 3f, g). Next, we sought to inhibit the activity of CRH neurons by using a recombinant AAV that conditionally expresses the inhibitory receptor hM4D(Gi) (Fig. 3b, inhibition setup). When activated by the designer drug clozapine-*N*-oxide (CNO), hM4D(Gi) suppresses neuronal firing¹². Inhibition of CRH-IRES-Cre neurons in the CeA and PVN was

verified by patch-clamp recording with brain slices (Extended Data Fig. 7d). Administration of CNO via the drinking water throughout the course of immunization significantly reduced the abundance of SPPCs in CRH-IRES-Cre mice as compared to control B6 mice (Fig. 3h, i). Together, these data show that basal CRH neuronal activity in the CeA and PVN are required for optimal SPPC formation.

To determine whether enhanced CRH neuronal activity would increase SPPC production during an immune response, we used a recombinant AAV that conditionally expresses the hM3D(Gq) receptor; as control, we used an AAV expressing a fluorescent protein (Fig. 3b, activation and control setup). Upon CNO activation, the hM3D(Gq) receptor triggers action potentials¹², as verified by recordings of single CRH neurons in brain slices (Extended Data Fig. 7e). As shown in Fig. 3j, k, when CNO was given from day 8 to day 12 after immunization, mice that harboured hM3D(Gq)-expressing CRH neurons produced more SPPCs than did control mice. Ablation, inhibition, or activation of CRH neurons did not significantly change the GC response (Fig. 3g, i, k).

To confirm that the connection between CeA and PVN CRH neurons and the splenic nerve is essential for optimal SPPC production, we denervated the spleen and then pharmacogenetically activated CeA and PVN CRH neurons during immunization. The two sets of surgery required—splenic denervation and transcranial injections (Extended Data Fig. 8a)—tended to cause variably high background levels of GCs and plasma cells (X.Z. et al., unpublished observations). We thus assessed NP-binding, isotype-switched IgG⁺ plasma cells, which were exclusively generated from the NP-KLH immunization. Pharmacogenetic activation of CRH neurons in the CeA and PVN led to an increase in NP-specific IgG⁺ plasma cells only in mice with intact splenic nerves, but not in denervated mice (Extended Data Fig. 8b, c).

Having identified a functionally important connection between CeA and PVN CRH neurons and the splenic nerve, we next investigated whether bodily behaviours could activate this pathway and thereby enhance the outcome of immunization. PVN CRH neurons respond to stress and drive the production of immunosuppressive glucocorticoids from the adrenal gland. For this reason, we thought that an immunostimulatory behavioural paradigm should provoke PVN CRH neurons but not too strongly, perhaps in the form of mild stress. We developed a protocol in which mice are made to stand on a round, transparent platform that is 10 cm in diameter and elevated to 1.5 m above the ground (Extended Data Fig. 9a). We call this behavioural regimen elevated-platform standing (EPS), during which mice exhibit signs of acrophobic stress (Supplementary Video 1). In CRH-IRES-Cre mice that had been stereotactically injected with a recombinant AAV that conditionally expresses the DIO-GCaMP6m calcium indicator, fibre photometry revealed that EPS induced calcium fluxes in CRH neurons of the CeA and PVN (Extended Data Fig. 9b, Fig. 4a, d). B6 mice were subjected to 3-min EPS twice daily between days 8 and 12 after NP-KLH immunization; whereas the GC magnitude was not changed compared to mice that had not undergone EPS, SPPC abundance was significantly increased (Fig. 4c). Notably, when spleen-denervated mice were immunized and subjected to EPS, the SPPC-enhancing effect of EPS was absent (Fig. 4d), indicating that the splenic nerve-dependent neural pathway is responsible for the EPS-mediated immunostimulatory effect.

To test whether all stressors would stimulate SPPC formation, we conducted similar analyses of mice subjected to prolonged physical restraint (PPR), a strong stress-inducing paradigm. As shown in Extended Data Fig. 9c, d, during a typical 90-min PPR session, PVN CRH neurons were strongly activated while CeA CRH neurons were suppressed. PPR led to an increase in circulating corticosterone levels as compared to EPS (Extended Data Fig. 9e, f). Notably, PPR during the course of immunization markedly suppressed GC formation and did not enhance SPPC formation (Extended Data Fig. 9g). Thus, only behavioural stress of appropriate form and strength, as represented by EPS, could be immunostimulatory.

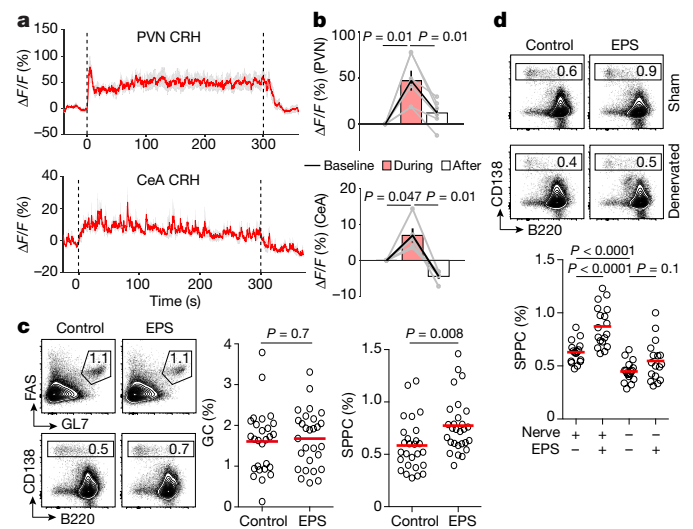


Fig. 4 | EPS stimulates CRH neurons and enhances plasma cell formation.

a, Representative traces of integrated calcium signals from PVN (top) or CeA (bottom) CRH neurons, read by photometry in CRH-IRES-Cre mice during EPS, presented as normalized changes in the GCaMP6m fluorescence intensity (mean in red, s.e.m. in grey). Dashed lines mark the beginning and the end of a 5-min EPS session. **b**, Average GCaMP6m signals, collected as in **a**, before, during and after a 5-min EPS session; grey lines, individual mice; black line with error bars, mean \pm s.e.m. of six mice. One-way ANOVA with Bonferroni's correction. **c**, Representative contour plots (left) and summary statistics of percentage GC and SPPC (right), 13 days after immunization. Data pooled from five independent experiments; each symbol indicates one mouse, lines denote means. **d**, Representative contour plots (top) and summary statistics of percentage SPPC (bottom), 13 days after immunization of B6 mice that were sham-operated or denervated in postnatal week 3 and allowed to recover for 6 weeks before immunization and EPS. Data pooled from three independent experiments; each symbol indicates one mouse, lines denote means. Two-tailed unpaired *t*-test (**c**, **d**).

To evaluate whether EPS-promoted SPPC formation is translated into an enhanced antigen-specific antibody response, we also measured serum NP-specific IgG titres 2 weeks after immunization. In EPS-experienced mice, titres increased by about 70% when compared with control mice (Fig. 5a). This EPS-induced increase required intact CRH neurons in the CeA and PVN, because no increase was detected when CRH neurons were ablated (Fig. 5b, c). The enhancing effect also depended on the splenic nerve, because it was not observed in spleen-denervated mice (Fig. 5d, e). The EPS effect also required expression of $\alpha 9$ AChRs on B cells (Fig. 5f, g). Finally, in a separate replication study, we followed the course of immunization for 4 weeks and found that, in addition to increased antibody titres (Extended Data Fig. 10a, b), EPS induced an increase in hypermutation-laden cells among SPPCs (Extended Data Fig. 10c), and led to a significant increase in the abundance of antigen-specific bone-marrow plasma cells (Extended Data Fig. 10d, e), which represent the long-lived compartment of humoral memory¹³. Together, these results indicate that, by increasing neural activity in the CeA/PVN–splenic nerve axis, bodily behaviours can enhance the outcome of immunization.

Activation of dopaminergic neurons in the ventral tegmental area (VTA) has been reported to enhance immunity¹⁴. Although chemo-genetic activation of neurons in the VTA did not increase GC or SPPC formation (data not shown), it is interesting to consider whether VTA dopaminergic neurons communicate with the CeA/PVN–splenic nerve pathway. A reflexive form of neuronal regulation of innate inflammation, the anti-inflammatory reflex, is well established¹⁵. ChAT-expressing T cells are essential for this anti-inflammatory reflex¹ and are probably involved in neural enhancement of the adaptive response described

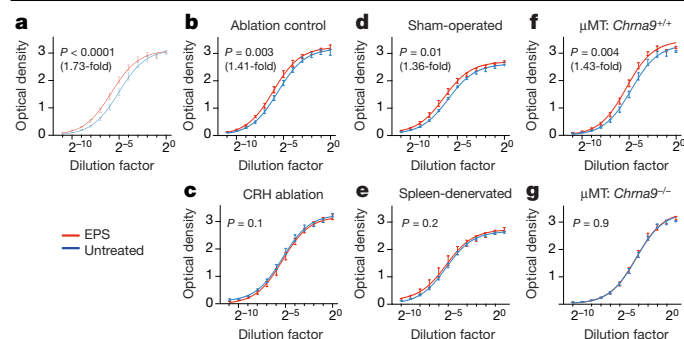


Fig. 5 | EPS enhances the antigen-specific antibody response via the PVN/CeA-splenic nerve axis. NP-specific IgG titres of NP-KLH-immunized control or EPS-experienced mice following manipulations of the CeA/PVN-splenic nerve axis. **a**, B6 mice that were untreated ($n = 17$) or subjected to EPS ($n = 16$). **b**, **c**, Control (**b**) or CeA/PVN CRH neuron-ablated (**c**) mice; untreated control ($n = 14$), EPS control ($n = 13$), untreated ablation ($n = 14$), EPS ablation ($n = 11$). **d**, **e**, Sham-operated (**d**) or spleen-denervated (**e**) mice; untreated sham-operated ($n = 5$), EPS sham-operated ($n = 6$), untreated denervated ($n = 6$), EPS denervated ($n = 6$). **f**, **g**, μ MT *Chrna9*^{+/+} (**d**) or μ MT *Chrna9*^{-/-} (**e**) chimaeric mice; untreated μ MT *Chrna9*^{+/+} ($n = 16$), EPS μ MT *Chrna9*^{+/+} ($n = 15$), untreated μ MT *Chrna9*^{-/-} ($n = 16$), EPS μ MT *Chrna9*^{-/-} ($n = 14$). On x-axis (dilution factor), 2^0 corresponds to 1:800 of the original serum. Mean \pm s.e.m. of all animals at each dilution overlaid with dose-response regression curves (goodness of fit $R^2 > 0.9$ in all cases). For each untreated-versus-EPS comparison, fold changes in IgG titres were calculated by EC_{50} of the respective curves, with between-group P values by extra sum-of-squares F -test.

here, even though our T-cell ablation study does not directly prove that such T cells relay noradrenergic signals to B cells.

Our surgical ablation of the splenic nerve, instead of catecholaminergic denervation by neurotoxins, allowed us to demonstrate that an efferent pathway that descends into the spleen is the mechanism responsible for the immunostimulatory effects of CRH neurons. These effects do not appear to involve acute reflexes, but rather hinge on tonic and provoked activities in CeA and PVN CRH neurons, which are intrinsically rhythmic¹⁶ and may be modulated by inputs from other brain areas¹⁷. The CRH-controlled HPA axis can also potently suppresses immune responses by releasing glucocorticoids. It is likely that, above a certain threshold of activation intensity, secretion of the HPA-driving hormone and consequent immunosuppression would dominate over the immunostimulatory effects of CRH neurons; below such a threshold, immunostimulation by neuronal signals descending into the spleen are more dominant.

In summary, our findings demonstrate brain control of adaptive immunity via direct neural connection. We speculate that bodily

behaviours and psychological conditioning could be used to induce stress of an appropriate strength and form, eventually to be defined in qualitative and quantitative terms of neuron activation in specific brain areas, in order to enhance host adaptive immunity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2235-7>.

- Rosas-Ballina, M. et al. Acetylcholine-synthesizing T cells relay neural signals in a vagus nerve circuit. *Science* **334**, 98–101 (2011).
- Rinner, I. & Schauenstein, K. Detection of choline-acetyltransferase activity in lymphocytes. *J. Neurosci. Res.* **35**, 188–191 (1993).
- Sato, K. Z. et al. Diversity of mRNA expression for muscarinic acetylcholine receptor subtypes and neuronal nicotinic acetylcholine receptor subunits in human mononuclear leukocytes and leukemic cell lines. *Neurosci. Lett.* **266**, 17–20 (1999).
- Wessler, I. & Kirkpatrick, C. J. Acetylcholine beyond neurons: the non-neuronal cholinergic system in humans. *Br. J. Pharmacol.* **154**, 1558–1571 (2008).
- Saalmüller, A. & Mettenleiter, T. C. Rapid identification and quantitation of cells infected by recombinant herpesvirus (pseudorabies virus) using a fluorescence-based β -galactosidase assay and flow cytometry. *J. Virol. Methods* **44**, 99–108 (1993).
- Cano, G., Sved, A. F., Rinaman, L., Rabin, B. S. & Card, J. P. Characterization of the central nervous system innervation of the rat spleen using viral transneuronal tracing. *J. Comp. Neurol.* **439**, 1–18 (2001).
- Keifer, O. P. Jr, Hurt, R. C., Ressler, K. J. & Marvar, P. J. The physiology of fear: reconceptualizing the role of the central amygdala in fear learning. *Physiology* **30**, 389–401 (2015).
- Herman, J. P. & Tasker, J. G. Paraventricular hypothalamic mechanisms of chronic stress adaptation. *Front. Endocrinol.* **7**, 137 (2016).
- Peng, J. et al. A quantitative analysis of the distribution of CRH neurons in whole mouse brain. *Front. Neuroanat.* **11**, 63 (2017).
- Kadmiel, M. & Cidlowski, J. A. Glucocorticoid receptor signaling in health and disease. *Trends Pharmacol. Sci.* **34**, 518–530 (2013).
- Vandevyver, S., Dejager, L., Tuckermann, J. & Libert, C. New insights into the anti-inflammatory mechanisms of glucocorticoids: an emerging role for glucocorticoid-receptor-mediated transactivation. *Endocrinology* **154**, 993–1007 (2013).
- Roth, B. L. DREADDs for neuroscientists. *Neuron* **89**, 683–694 (2016).
- Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* **15**, 160–171 (2015).
- Ben-Shaanan, T. L. et al. Activation of the reward system boosts innate and adaptive immunity. *Nat. Med.* **22**, 940–944 (2016).
- Tracey, K. J. The inflammatory reflex. *Nature* **420**, 853–859 (2002).
- Lightman, S. L. & Conway-Campbell, B. L. The crucial role of pulsatile activity of the HPA axis for continuous dynamic equilibration. *Nat. Rev. Neurosci.* **11**, 710–718 (2010).
- Ulrich-Lai, Y. M. & Herman, J. P. Neural regulation of endocrine and autonomic stress responses. *Nat. Rev. Neurosci.* **10**, 397–409 (2009).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Mice

C57BL/6J (Jax 664), CD45.1 (Jax 2014), μ MT (Jax 2288), *Tcrb*^{-/-} *Tcrd*^{-/-} (Jax 2122), ChAT-IRES-Cre (Jax 31661), CRH-IRES-Cre (Jax 12704), Rosa26-iDTR (Jax 7900), Rosa26-Ai3 (Jax 7903) and Rosa26-Ai14 (7914) mice were originally from the Jackson Laboratory and maintained on a B6 background. Relevant mice were interbred to obtain CRH-IRES-Cre:Rosa-Ai3, ChAT-IRES-Cre:Rosa26-iDTR and ChAT-IRES-Cre: Ai14 mice. All mice were housed as groups of 4 to 6 individuals per cage and maintained on a 12-h light-dark cycle at 22–25°C under specific-pathogen free conditions. All animal experiments were approved by the Institutional Animal Care and Use Committee and conducted in accordance of governmental and Tsinghua guidelines for animal welfare. When relevant and applicable, age- and sex-matched mice were randomly chosen from same cages to be included in experimental and control groups. The investigators were not blinded to allocation during experiments and outcome assessment.

Construction of AChR-deficient mice

For germline ablation of the *Chrna9*, *Chrn4*, and *Chrn1* genes in B6 mice, the standard CRISPR–Cas9 technique was used. In brief, for each gene, two sgRNAs targeting sequences coding for the transmembrane domain were co-injected together with Cas9-coding mRNA into fertilized B6 eggs. Offspring were genotyped by sequencing to identify mutants that cannot normally express the transmembrane domain owing to either deletion of the coding sequence or introduction of premature stop codons. The chosen mutant founders were then backcrossed with B6 mice for at least three generations before subsequent experiments. The sgRNA sequences used were: *Chrna9* GCGGAGCGAGGAACGATATG and GAGATGACGTTCTTCACGGC; *Chrn4* TCACTGTCCTTCGACCCGGG and TTGTGTGCTCATCAGTCGC; *Chrn1* ATGCATCCTCATCAGCTCC and CTAGGTTAATCCAATTCCA. The genotyping primers used were: 5'-GCAGTGCAACCTGACC TTTG (*Chrna9*-F), 5'-ACGCCATCACAAGTCTACA (*Chrna9*-R), 5'-TTCCCTTCGACGACGAGAAC (*Chrn4*-F), CACACAGTGGTGACG ATGGA (*Chrn4*-R). Also see Extended Data Fig. 3 for details.

Construction of mixed bone-marrow chimaeras

Sex-matched B6 recipient mice were lethally irradiated by X-ray (6 Gy, twice) and then intravenously infused with a combination of 4×10^6 bone-marrow cells of the indicated genotypes. Chimeras were given antibiotics in drinking water for the first 2 weeks and used for experiments 8 weeks after reconstitution.

Adoptive transfer and reconstitution of mice deficient in T or B cells

CD4⁺ T cells and CD19⁺ B cells were isolated using a CD4 T cell isolation kit and CD19 Microbeads (Miltenyi Biotec), respectively, according to the manufacturer's protocols. A total of 4×10^6 CD4⁺ T cells or 1.5×10^7 B cells of the indicated genotypes were intravenously infused into *Tcrb*^{-/-} *Tcrd*^{-/-} mice or μ MT mice, respectively, to create animals with a CD4 or B cell compartment of the desired genotype. Five days after the transfer, mice were used for experiments as indicated.

Splenic denervation

Three-week-old male B6 mice were anaesthetized with 75 mg/kg sodium pentobarbital. The peritoneal cavity was accessed through a midline abdominal incision. Forceps were used to blunt-isolate the spleen away from the peritoneal cavity so that the three main supplying vasculature trees were clearly exposed (Extended Data Fig. 1a). With the peritoneal cavity and other organs protected with a moistened cotton pad (Extended Data Fig. 1a), under a dissection microscope, absolute ethanol was repeatedly applied with cotton tips to those vasculature trees for 5–10 s each time, at 5-s intervals, and

about seven times in total in order to deplete the splenic nerve fibres that run along them (Extended Data Fig. 1c). Care was taken to avoid excessive ethanol dripping from the cotton tip and to avoid causing visible vessel spasms, which could lead to permanent damage to blood vessels and cause splenic necrosis and complete organ absorption. For sham-operated mice, the entire surgical operation was identical except that saline (pH 7.4) instead of absolute ethanol was repeatedly applied. After denervation surgery, animals were allowed to recover for 6 weeks before immunization and other experiments. Disappearance of splenic innervation was histologically verifiable a week after the surgery (Extended Data Fig. 1f, g).

Immunization

To measure adaptive immune responses, mice of indicated types and treatments were intraperitoneally injected with 100 μ g NP-KLH or 60 μ g NP-Ficoll (Biosearch Technologies) mixed with 1 μ g LPS (Sigma) in alum (Thermo Scientific).

Acetylcholine and norepinephrine treatment

For certain experiments, noradrenaline bitartrate monohydrate or acetylcholine chloride (Sigma) or the vehicle phosphate buffer saline (pH 7.4) was injected subcutaneously in a volume of 100 μ l twice daily, 10 h apart, at indicated concentrations during days 8–12 after immunization.

Retroviral transduction of B cells and neurotransmitter tests

The coding sequence of mouse AchE (NCBI: 11423) or COMT (NCBI: 12846) was inserted into multiple clone sites of a pBMN-PIB retroviral vector, upstream of an internal ribosomal entry site (IRES) and enhanced green fluorescent protein. Cultures of Plat-E packaging cell line were transfected with these constructs, and supernatants containing retrovirus were collected for transduction. B cells were isolated using CD19 Microbeads (Miltenyi Biotec) and activated with 1 μ g/ml LPS (Sigma) for 1 day before being spin-infected with retroviral supernatants at 1,500g for 2 h, as previously described¹⁸. A total of 4×10^6 AchE- or COMT-transduced CD45.1B cells was intravenously transferred into B6 mice that had already been immunized with NP-KLH, once on day 7 and one more time on day 9 after immunization.

Flow cytometry

Single-cell suspensions of spleen or bone marrow were incubated in MACS buffer (PBS supplemented with 1% FBS and 5 mM EDTA) containing 20 μ g ml⁻¹ 2.4G2 antibody (BioXcell) for 20 min and then stained with the indicated antibodies. Staining reagents included APC-Cy7 anti-CD19 (1D3), AF700 or BV421 anti-CD4 (GK1.5), PE anti-CD8 (53-6.7), AF700 anti-CD3 (17A2), AF700 anti-CD11c (HL3), biotin-CD43 (S7), PE anti-CD19 (6D5), AF700 or APC-cy7 anti-B220 (RA3-6B2), PE-Cy7 anti-CD93 (A4.1), FITC anti-CD23 (B3B4), APC anti-CD21 (7E9), PE-Cy7 anti-Fas (Jo2), FITC or APC anti-GL7 (GL-7), BV510 or PE or PE-Cy7 anti-CD138 (281-2), EF450 anti-IgM (EB121-15F9), APC and FITC anti-IgM (II/41), BV421 anti-IgM (eB121-15F9), percpCy5.5 anti-IgD (11-26c.2a), EF450 anti-IgD (11-26C), FITC anti-IgG (poly4053), Pacific Blue anti-CD45.1 (A20), APC-Cy7 anti-CD45.2 (104), V450 anti-Gr1 (RB6-8C5), APC streptavidin (Biolegend) and EF450 Streptavidin (eBioscience). NP-PE (Biosearch Technologies) and Fixation/Permeabilization kit (554714, BD) were used for intracellular staining of NP-binding plasma cells. CaspGLOW Fluorescein Active Caspase Staining Kit (K180-100, Biovision) was used to detect apoptosis of CD4⁺ and CD19⁺ cells ex vivo. Cells were typically stained on ice with primary reagents for 30–60 min followed by staining with secondary reagents for 30 min. Data were collected on an LSR II cytometer (BD) and analysed with FlowJo software (TreeStar). Where applicable, cell sorting was conducted with Aria IV (BD). Dead cells and cell aggregates were excluded from analyses by 7-AAD staining (Biotium) and FSC-H/FSC-A characteristics.

Immunohistochemistry

For spleen immunohistochemistry, all samples were fixed in 4% paraformaldehyde, embedded in OCT compound (Tissue-Tek), and kept at -20°C until further processing. Spleen slices (40 to 50 μm in thickness) were cut using a microtome (Leica), blocked with PBS containing 0.3% Triton X-100 and 3% FBS for 30 min at room temperature, and then incubated with primary antibodies at appropriate final dilutions in a humidified chamber at 4°C overnight. Primary antibodies included anti-tyrosine hydroxylase (AB152, Millipore; 1:400), EF450 anti-IgD (1:100), EF660 anti-CD3 ϵ (17A2, BD; 1:100), and goat-anti-Igk (1050-01, Southern Biotech; 1:400). After staining with primary antibodies, the tissue slices were washed three times in washing buffer (PBS, 0.3% Triton X-100) and incubated with AF488 donkey anti-rabbit (A21206, Life Technology; 1:400) and AF647 donkey anti-goat (A21447, Life Technology; 1:400) for 30 min. After washing, the tissue slices were dried, mounted with ProlongGold (Invitrogen) and examined using a Nikon A1Rsi confocal microscope. Images were analysed using the Imaris software package (Bitplane).

Measurement of splenic norepinephrine levels by ELISA

Spleens of sham-operated or denervated mice were weighed and then cut into small pieces of approximately 1 mm^3 . Tissue fragments were incubated in ice-cold PBS containing 1% EDTA and shaken at 100 rpm in a 35-mm dish on a shaker for 15 min. Norepinephrine in supernatants was then measured using the Norepinephrine ELISA kit (KA1877, Abnova) according to the manufacturer's instruction, and the final results were presented as spleen weight-normalized norepinephrine mass concentration.

Measurement of serum NP-specific IgG by ELISA

Sera were harvested from relevant mice and frozen until testing. A 96-well ELISA plate (42592, Costar) was coated with NP₂₃-BSA (Biosearch Technologies) in PBS overnight and washed. A serial 1:2 dilution of each serum sample, starting from 1:800, was then loaded into the plate and incubated at 37°C for 1 h. After washings, the plate was further incubated with HRP-conjugated goat anti-mouse IgG (poly4053, Biolegend) at 1:20,000 for 1 h before development with TMB Substrate Set (421101, Biolegend). The chromogenic reaction was stopped with 1M HCL, and optical density (OD) was read on an iMark plate reader. For each group of animals, the resulting dilution curves were fit with the 3-parameter dose-response curve in Prism (Graphpad). Curves from two groups in comparison were analysed by the extra sum-of-squares *F*-test, and their nominal EC₅₀s were used to calculate fold-change in NP-specific IgG titres.

Synthesis of ACh analogue and detection of surface ACh binding

Our choice of the synthetic acetylcholine (ACh) analogue for detecting acetylcholine receptors (2-(2-azidoacetoxy)-N,N,N-trimethylethan-1-aminium bromide, see Extended Data Fig. 3g for reaction scheme) was based on the structure of the agonist-M2 receptor complex¹⁹. In brief, to a solution of 2-azidoacetic acid I (500 mg, 1.0 equiv, 4.95 mmol) in dichloromethane (7 ml) was added a few drops of DMF under argon. After the addition of oxalyl dichloride (796 μl , 1.9 equiv, 9.41 mmol) was complete at 0°C , the reaction mixture was stirred for a further 5 min at the same temperature. Then the reaction mixture was warmed to room temperature and stirred for another 3 h. After the solvent and other volatile components had evaporated, the resulting material was dissolved in dichloromethane (7 ml). To this mixture was added 2-bromoethan-1-ol (182 μl , 0.52 equiv, 2.57 mmol) and pyridine (588 μl , 1.48 equiv, 1.41 mmol) at 0°C . After being stirred at 0°C for 5 min, the reaction mixture was warmed to room temperature and stirred for another 1 h. Then the organic layer was separated, and the aqueous solution was extracted with dichloromethane. The combined organic layer was washed with saturated aq. NaHCO₃ brine, dried over Na₂SO₄

and concentrated in vacuo. The crude residue was purified through silica gel chromatography (1:10 ether:hexanes) to give 201.5 mg (38%) of II as oil. To a 25 ml flame-dried round flask was added c (80 mg, 1.0 equiv, 0.38 mmol) and toluene (5 ml). Then a solution of trimethylamine in ethanol (1.2 ml, 3.2 M) was added by syringe. After being stirred at 90°C for 5 h, the reaction mixture was cooled to room temperature. The precipitate was collected by filtration and washed with toluene and dichloromethane to obtain 44.3 mg (43%) of III as solid. ¹H NMR (400 MHz, CD₃OD) δ 4.70–4.62 (m, 2 H), 4.01 (s, 2 H), 3.82–4.76 (m, 2 H), 3.24 (s, 9 H); ¹³C NMR (100 MHz, CD₃OD) δ 169.5, 65.9, 59.9, 54.5, 51.1; MS(ESI) *m/z* calculated for [M-Br]⁺ 187.12, found 187.25.

To detect acetylcholine-binding cells, a single-cell suspension was blocked in MACS buffer containing 20 $\mu\text{g ml}^{-1}$ 2.4G2 antibody (BioXcell) for 20 min and then incubated with the chemical III (1:200) on ice for 1.5 h. After being washed four times, these cells were incubated at room temperature with 50 μM of DBCO-biotin in PBS containing 1% FBS in the dark for 30 min. After being washed four times, the cells were stained on ice for 30 min with 1 $\mu\text{g/ml}$ EF450 streptavidin (eBioscience) together with other antibodies for surface phenotyping.

Survey detection of AChRs by RT-PCR

Total RNA was extracted from sorted splenic CD138⁺B220^{low} SPPCs, total B cells or Fas⁺GL7⁺ GC cells using the RNeasy PLUS Mini/Micro Kit (Qiagen) and cDNA was prepared using RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher). The primers used for various receptors (Invitrogen) are listed in Supplementary Table 3. Bands of expected sizes from electrophoresis gels were used to retrieve DNA for sequencing verification of the amplicon identities.

Quantitative real-time RT-PCR analysis for ChAT expression in T cells

CD4 T cells from ChAT-IRES-Cre; Rosa26-Ai14 mice were transferred into *Tcrb*^{-/-}*Tcrd*^{-/-} mice and immunized as described above. On day 8, CD4 T cells from these mice were isolated using a CD4 T cell isolation kit (Miltenyi Biotec) according to the manufacturer's protocol. These cells were further sorted into ChAT-Ai14⁺ and ChAT-Ai14⁻ fractions. Sorted ChAT-Ai14⁺ and ChAT-Ai14⁻ CD4 T cells were lysed in hypotonic lysis buffer (Amresco, M334) at 72°C for 3 min. Smart-seq2 reverse transcription was then conducted. After pre-amplification and purification with AMPure XP beads, cDNA was used for qPCR. qPCR was performed using TB Green Premix EX Taq (RR420A, TaKaRa), and primer sequences were as follows: *Chat* forward 5'-AGGGCAGCCTCTCTGTATGA-3'; *Chat* reverse 5'-ATCCTCGTTGGACGCCATTT-3'; *Gapdh* forward 5'-CAAGGTCATCCATGACAACCTTG-3'; *Gapdh* reverse 5'-GTCCACCACCTGTTGCTGTAG-3'.

Mutation analysis of V_H186.2-carrying SPPCs

To analyse the heavy-chain sequences of NP-specific SPPCs, CD138⁺ cells were isolated 13 days after NP-KLH immunization. After incubation of cells in lysis buffer at 60°C for 5 min, reverse transcription was carried out using Superscript cDNA Synthesis Kit (Invitrogen) according to the manufacturer's protocol. V_H186.2 was amplified by nested PCR with the following primers (first primer: 5'-CTCTTCTTGGCAGCAACAGC-3', first antisense primer: 5'-GCTGCTCAGAGTGTAGAGGTC-3', second primer: 5'-GTGTCCACTCCAGCTCCAAC-3', second antisense primer: 5'-GTTCCAGGTCAGTGTCACTG-3'). PCR products (400 bp) were purified by gel electrophoresis, cloned into a T vector (Takara) and sequenced. Mutations were identified by comparing these sequences to the germline V_H186.2 sequence, and only non-identical sequences were counted as clones.

Stereotactic viral microinjection to PVN and CeA

AAV vectors containing the DIO-taCaspase3 and DIO-GCaMP6m constructs were packaged into an AAV2/9 serotype to achieve the indicated infectious units per ml in the laboratory of Minmin Luo

(National Institute of Biological Sciences, Beijing, China). DIO-eYFP, DIO-hM3D(Gq)-mCherry, and DIO-hM4D(Gi)-mCherry were from Vigenebio Inc. Surgical procedures were as previously described²⁰. In brief, mice were anaesthetized by intraperitoneal injection of pentobarbital (100 mg kg⁻¹), kept warmed with an electric heating pad (BrainKing Biotech), and mounted on a stereotaxic apparatus to adjust the skull in parallel to the reference panel. Using a microsyringe pump (Nanject III #3-000-207, DRUMMOND), AAV virus was slowly injected (20 nl min⁻¹) into PVN (Bregma coordinates: -0.6 mm AP, ±0.2 mm ML, -4.8 mm DV) and CeA (Bregma coordinates: -1.23 mm AP, ±2.75 mm ML, -4.5 mm DV), unilaterally for fibre photometry or bilaterally for pharmacogenetic experiments. The pump was maintained in position for an additional period of 10 min before slow withdrawal to allow viral particles to diffuse and be absorbed at the injection site. For fibre photometry, coordinates for fibre implantation were offset relative to the viral injection sites (Bregma for PVN: -0.6 mm AP, ±0.2 mm ML, -4.5 mm DV; bregma for CeA: -1.23 mm AP, ±2.75 mm ML, -3.9 mm DV). Specific volumes of AAV injected to each site were as follow: AAV2/9-CAG-DIO-taCasp3-TEVp (1.72×10^{13} IFU/ml; 100 nl to PVN and CeA), AAV2/9-hsyn-DIO-GCaMP6m ($1-5 \times 10^{12}$ IFU/ml; 300 nl to PVN and CeA), AAV9-EF1α-DIO-hM3D(Gq)-mCherry (1.85×10^{13} vector genomes (v.g.)/ml; 120 nl to PVN and 150 nl to CeA), AAV9-EF1α-DIO-hM4D(Gi)-mCherry (1.53×10^{13} v.g./ml; 120 nl to PVN and 150 nl to CeA), AAV9-EF1α-DIO-eYFP (5.23×10^{13} v.g./ml; 120 nl to PVN and 150 nl to CeA), AAV2/9-EF1α-DIO-ChR2-mCherry ($1-5 \times 10^{12}$ v.g./ml; 300 nl to PVN and 300 nl to CeA).

PRV retrograde tracing from the spleen

Pseudorabies virus (PRV-CMV-mRFP or PRV-CMV-GFP) was prepared as stock of 10^9 genomic copies per ml. Adult B6 mice were anaesthetized with 75 mg/kg sodium pentobarbital. The spleen was accessed through a midline opening to the peritoneal cavity. One microlitre of the viral stock solution was injected into each of the upper and lower tips of the spleen using a syringe fitted with a needle 30 µm in diameter. The surgical wound was closed by standard sutures and the indicated numbers of mice were killed 24, 48, 72 and 96 h later to retrieve the spinal cord and brain. Sections were prepared and processed for identification of PRV-mRFP⁺ neurons in different regions. In brief, brains were fixed in 4% PFA overnight and dehydrated by 30% sucrose at 4 °C for 2–3 days. Coronal sections (35 µm) were cut with a cryostat microtome (Leica). Tissue sections were blocked with PBS containing 5% BSA and 0.3% Triton X-100, stained with rabbit polyclonal anti-dsRed (632496, Clontech; 1:1,000) for 48 h at 4 °C, washed three times, and then stained with secondary antibody AF594 donkey anti-rabbit IgG (1:1,000) for 2 h at room temperature.

Images were then acquired with an Olympus VS120 microscope and analysed using Nikon AIR software. Different brain regions were identified using the Allen Mouse Brain Atlas (<https://mouse.brain-map.org/>) and the Mouse Brain in Stereotaxic Coordinates²¹.

Recording of splenic-nerve activity during optogenetic stimulation of CRH neurons

CRH-IRES-Cre mice were stereotactically infected with AAV2/9-EF1α-DIO-ChR2-mCherry at the CeA and PVN. Fibre implantation was offset relative to the viral injection sites (bregma for PVN: -0.6 mm AP, ±0.2 mm ML, -4.5 mm DV; bregma for CeA: -1.23 mm AP, ±2.75 mm ML, -3.9 mm DV). Two weeks later, these mice were anaesthetized by an intraperitoneal injection of 1% pentobarbital sodium. The splenic nerve was exposed in a similar manner as described for denervation surgery above. The electrode was custom-made by CorTec GmbH (1041.2406.51 - 2 Micro Cuff Sling 100/Pt-Ir/ 1mm long/0.35mm C2C/0.3 × 0.5mm Opening/ Cable entry top - welded) and was implanted around the splenic nerve so that the nerve could sit atop the metallic surface of the electrode probe. The implanted electrode was connected to a multichannel recording and signal processing system (LabChart, ADInstruments). To activate CRH neurons expressing ChR2 for each

trial, a 473-nm laser light was delivered as 2-ms pulses at 50 Hz for 200 ms every second. Light intensity was adjusted with an optical power meter (Sanwa) to reach 10 mW at the end of the implanted fibre stub. Recording signals were sampled at 1,000 Hz. For data analysis, all signals were digitally bandpass-filtered between 10 and 60 Hz to reduce noise. Firing spikes were automatically called by Matlab (R2016b) function *findpeaks* with a threshold set at 120 µV. Average firing rates during 20 s before and 20 s after optogenetic stimulation of CRH neurons were calculated and statistically analysed by paired *t*-test.

Pharmacogenetic studies

For activation of CRH neurons in PVN and CeA, groups of male CRH-IRES-Cre mice were stereotactically injected with DIO-hM3D(Gq)- or control AAV into the PVN and CeA regions. Under anaesthesia by 1.5% isoflurane, these mice were intraperitoneally injected with 1 mg/kg CNO (Enzo) twice daily, -10 h apart, between days 8 and 12 after NP-KLH immunization. For inhibition of CRH neurons in PVN and CeA, male CRH-IRES-Cre or control B6 mice were maintained with drinking water that was supplemented with CNO to achieve a daily intake of ~5 mg/kg CNO, assuming water consumption of 5 ml per day per mouse. The CNO-supplemented water was given from 2–3 days before immunization to 12 days after immunization.

The EPS behaviour regimen

Mice were individually placed on a transparent plastic platform of 10 cm in diameter and raised 150 cm above the ground. For each session, the animal was acoustically and visually isolated for the indicated amount of time. To test the effect on GC and plasma cell formation, mice were subjected to 3 min EPS twice daily for the indicated number of days, one administered between 10:00 and 12:00 and the other between 22:00 and 24:00 each day.

The PBR behaviour regimen

Mice were individually placed into a Mouse DecapiCone disposable restrainer (Braintree Scientific Inc.), with the open end of the restrainer cable-tied near the base of the mouse tail. For each session, the animal was acoustically and visually isolated for 90 min. To test the effect on GC and plasma cell formation, mice were subjected to the PBR twice daily as scheduled for EPS.

Fibre photometry

Following AAV-DIO-GCaMP6m or control AAV-DIO-GFP virus injection, an optical fibre with an outer diameter of 200 µm and 0.37 numerical aperture (AniLab) was placed in a ceramic ferrule and inserted towards the PVN through the craniotomy. Mice were individually housed for at least 2 weeks to recover. Fluorescence signals were acquired with a fibre photometric system equipped with a 488 nm excitation laser, 505–544 nm emission filter and a photomultiplier tube (R3896, Hamamatsu). The analogue voltage signals were digitalized at 100 Hz and recorded using a Power 1401 digitizer and Spike2 software (CED, Cambridge, UK). To connect the implanted fibre and the photometric system, an optical fibre (RJPSF2, Thorlabs) with integrated rotary joint was used to prevent fibre damage as a result of animal movement. The laser power was adjusted at the tip of the optical fibre to a low level of 20–40 µW in order to minimize bleaching. Acquired photometry data were exported to Matlab R2016b mat files from Spike2 for further analyses. Data were segmented according to EPS behavioural events within individual trials. The normalized change in fluorescence signal ($\Delta F/F$) was calculated as $(F - F_0)/F_0$, where F_0 is the average baseline fluorescent signal before EPS. $\Delta F/F$ values were plotted as an average trace bracketed by a shaded area indicating s.e.m.

Preparation of brain slices

Adult (8–12-week-old) male CRH-IRES-Cre mice: Rosa26-Ai3 mice were anaesthetized by intraperitoneal injection of pentobarbital (100 mg kg⁻¹) and then perfused transcardially with ice-cold, oxygenated (95% O₂/5% CO₂)

Article

NMDG artificial cerebrospinal fluid (ACSF) solution (93 mM NMDG, 93 mM HCl, 2.5 mM KCl, 1.25 mM NaH_2PO_4 , 10 mM $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 30 mM NaHCO_3 , 25 mM glucose, 20 mM HEPES, 5 mM sodium ascorbate, 3 mM sodium pyruvate, 2 mM thiourea). After perfusion, the brain was rapidly removed, immediately transferred into ice-cold, oxygenated NMDG ACSF solution, and sectioned coronally into 280- μm slices with a vibratome (VT1200 S, Leica). Brain slices containing the PVN and CeA were incubated in oxygenated NMDG ACSF solution at 32 °C for 15 min, and then transferred into a normal oxygenated ACSF solution (126 mM NaCl, 2.5 mM KCl, 1.25 mM NaH_2PO_4 , 2 mM $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 10 mM glucose, 26 mM NaHCO_3 , 2 mM CaCl_2) for incubation at room temperature for 1 h. All chemicals used in slice preparation were purchased from Sigma-Aldrich (St. Louis, MO, USA).

Electrophysiological recording from brain slices

Slices were placed in the recording chamber, submerged and perfused with ACSF at a rate of 3 ml min^{-1} at 28 °C. CRH⁺ positive neurons were identified by differential interference contrast optics (DIC; Olympus BX61WI). The recording pipettes (3–4 M Ω) were prepared using a micropipette puller (P97, Sutter Instrument; USA). To analyse firing rates, cell-attached or whole-cell recording was conducted. For cell-attached recording, the pipette was filled with ACSF solution. Cells were held at 0 pA under a current-clamp mode to record spontaneous firing. CNO (5 μM) was locally perfused to the cell through a drug perfusion system. For each recording, the baseline of spontaneous firing was recorded for at least 3 min before CNO was given. For whole-cell recording, the pipette was filled with ACSF solution containing 133 mM potassium gluconate, 18 mM NaCl, 0.6 mM EGTA, 10 mM HEPES, 2 mM Mg-ATP, and 0.3 mM Na_3GTP (pH 7.2, 280 mOsm). All recordings were acquired using a Multiclamp 700B amplifier and signals were low-pass filtered at 3 kHz and digitized at 10 kHz (DigiData 1550, Molecular Devices). Data were analysed using Clampfit 10 software (Molecular Devices), Mini Analysis Program (Synaptosoft) and Matlab R2016b programs.

Measurement of serum corticosterone

Mice were bled from the orbital sinus under pentobarbital anaesthesia. Blood samples were left to coagulate at room temperature for 2 h before serum harvesting by 10-min centrifugation at 3,000g. Sera were mixed with 100% methanol and centrifuged for 10 min at 4 °C. The supernatants, containing free-state corticosterone, were then vacuum-dried (Savant SpeedVac SPD121P, Thermo Fisher) and submitted for LC–MS. Detection and quantification of corticosterone were carried out using a Waters Acquity I Class UPLC system connected to an AB Sciex tripleTOF mass spectrometer with electrospray ionization. The compound was detected in positive ion mode. Calibration curves from 14.43 to 577.27 nmol/l were constructed using commercial corticosterone standard (CAS No. 50-22-6, Cayman) and the peak area was linear to concentration over this range ($R^2 > 0.99$). The area-under-the-curve upon LC–MS elution was used to quantify the corticosterone level in the samples.

Statistical data analysis

Statistics and graphing were conducted in Prism (Graphpad). Unless specifically noted otherwise, two-tailed unpaired Student's *t*-tests were used to compare the endpoint means of two groups. Non-parametric, two-tailed Mann–Whitney tests were used in cases of highly skewed distributions.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data generated here are included within the paper (and its Supplementary Information files) or are available from the corresponding authors upon reasonable request. Source data for Figs. 1–5 and Extended Data Figs. 1–10 are provided with the paper.

18. Wang, Y. et al. Germinal-center development of memory B cells driven by IL-9 from follicular helper T cells. *Nat. Immunol.* **18**, 921–930 (2017).
19. Kruse, A. C. et al. Activation and allosteric modulation of a muscarinic acetylcholine receptor. *Nature* **504**, 101–106 (2013).
20. Li, Y. et al. Serotonin neurons in the dorsal raphe nucleus encode reward signals. *Nat. Commun.* **7**, 10503 (2016).
21. Paxinos, G. & Franklin, K. B. J. *Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates* 4th edn (Academic, 2012).

Acknowledgements We thank S. Chavan for sharing technical information about electrophysiological recording. The work was funded in part by the National Key R&D Program of China (Ministry of Science and Technology, 2018YFE0200300 to H.Q.), National Natural Science Foundation of China (grants 81621002, 31830023 to H.Q.; grants 61890951, 61890950, 31671086 to J.H.), the Tsinghua-Peking Center for Life Sciences, and the Beijing Municipal Science & Technology Commission.

Author contributions H.Q. conceptualized and supervised the study. X.Z. developed the splenic denervation technique and, together with L.Z., the EPS regimen. X.Z., B.L. and Y.Y. participated in designing the overall study together with H.Q., J.H., and Y.Z. X.Z. and L.Z. conducted most of the immunological experiments. L.H. synthesized the acetylcholine analogue under the supervision of X.L. Y.Y., S.J., and L.Z. conducted the PRV tracing experiments under the supervision of F.X. and J.H. X.Z. and Y.Y. conducted electrophysiological recording of the splenic nerve. X.Z., Y.Y. and L.Z. conducted CRH neuron ablation and fibre photometry studies under the supervision of J.H., W.S. and H.Q. B.L., X.Z., B.K. and L.Z. conducted CRH neuron inhibition and activation studies under the supervision of Y.Z. and H.Q. All authors contributed to data interpretation. H.Q. and X.Z. wrote the paper with input from all authors.

Competing interests The authors declare no competing interests.

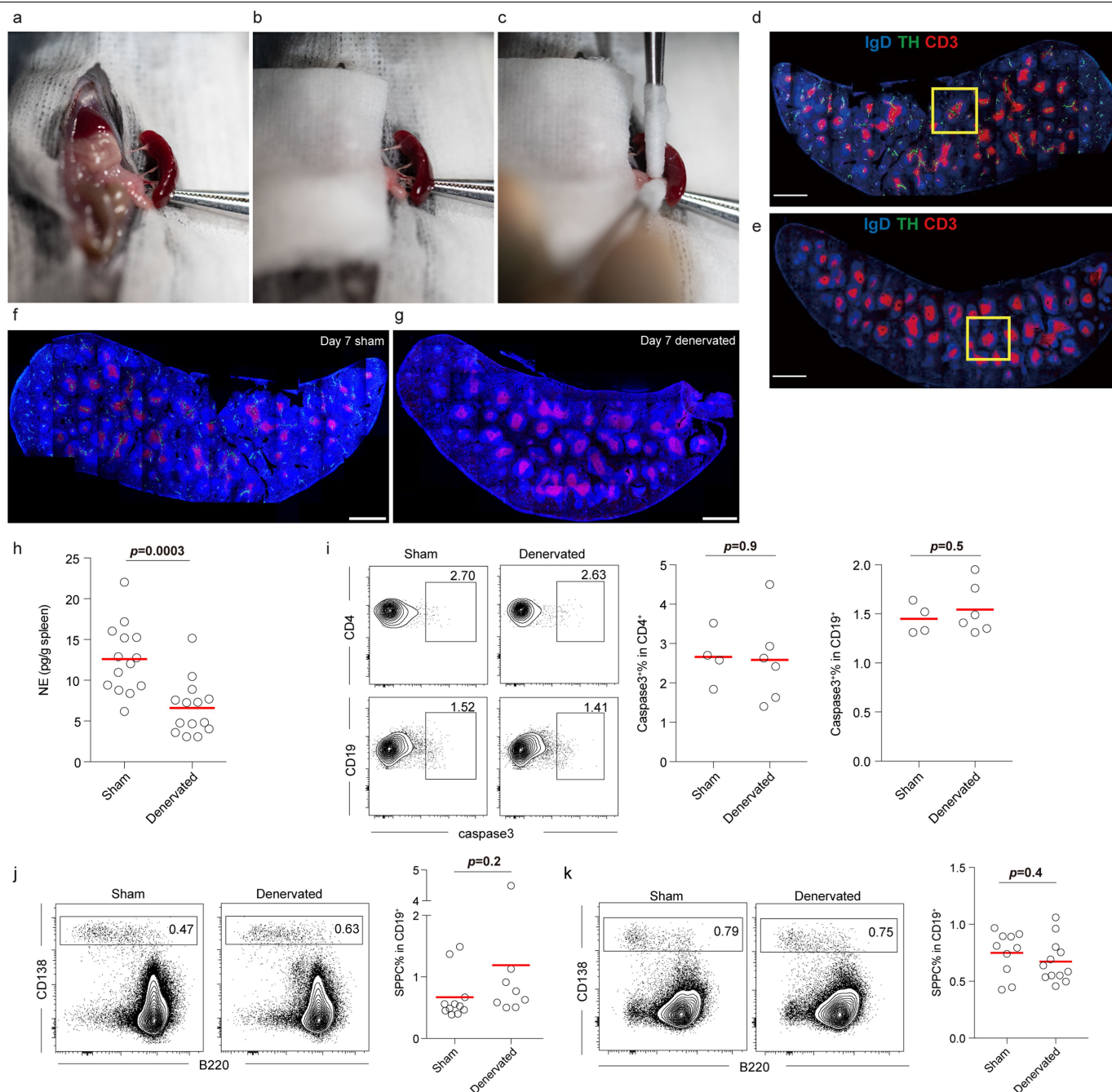
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2235-7>.

Correspondence and requests for materials should be addressed to Y.Z., J.H. or H.Q.

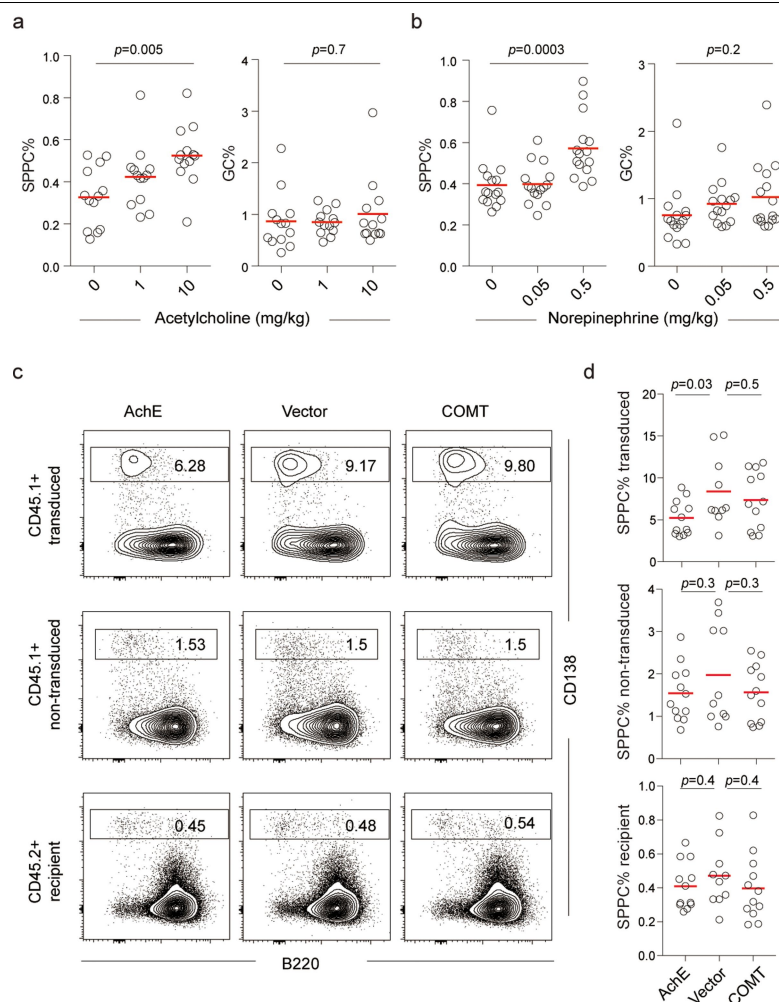
Peer review information Nature thanks Jonathan Kipnis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Surgical splenic denervation and its lack of effects on plasma cell formation in a T-independent response. **a–c**, Photographs of surgical exposure and alcohol application for denervation or sham operation. See Methods for detailed description. **d, e**, Whole spleen cross-sections from sham-operated (**d**) and denervated (**e**) mice, from which the images in Fig. 1a were cropped (yellow boxes). **f, g**, Representative images of whole spleen sections of sham-operated (**f**) or denervated (**g**) mice 1 week after surgery. Data representative of two experiments. Green, TH staining; blue, IgD staining of follicle; red, CD3ε staining of T-cell zone. **h**, Norepinephrine (NE) concentrations in splenic tissue from sham-operated and denervated mice 13

days after NP-KLH immunization. Data pooled from three experiments. Each symbol denotes one mouse, lines denote means. **i**, Representative contour plots (left) and summary apoptotic frequencies (right) in CD4 T and B cell compartments of sham and denervated mice. Each symbol denotes one mouse, lines denote means. **j, k**, Percentage SPPC 7 days (**j**) or 13 days (**k**) after NP-Ficoll immunization in sham and denervated mice. Representative FACS contour plots (left) and summary data in scatter plots (right). Data pooled from two independent experiments, with each symbol indicating one mouse; lines denote means. Two-tailed unpaired *t*-test (**h–k**).

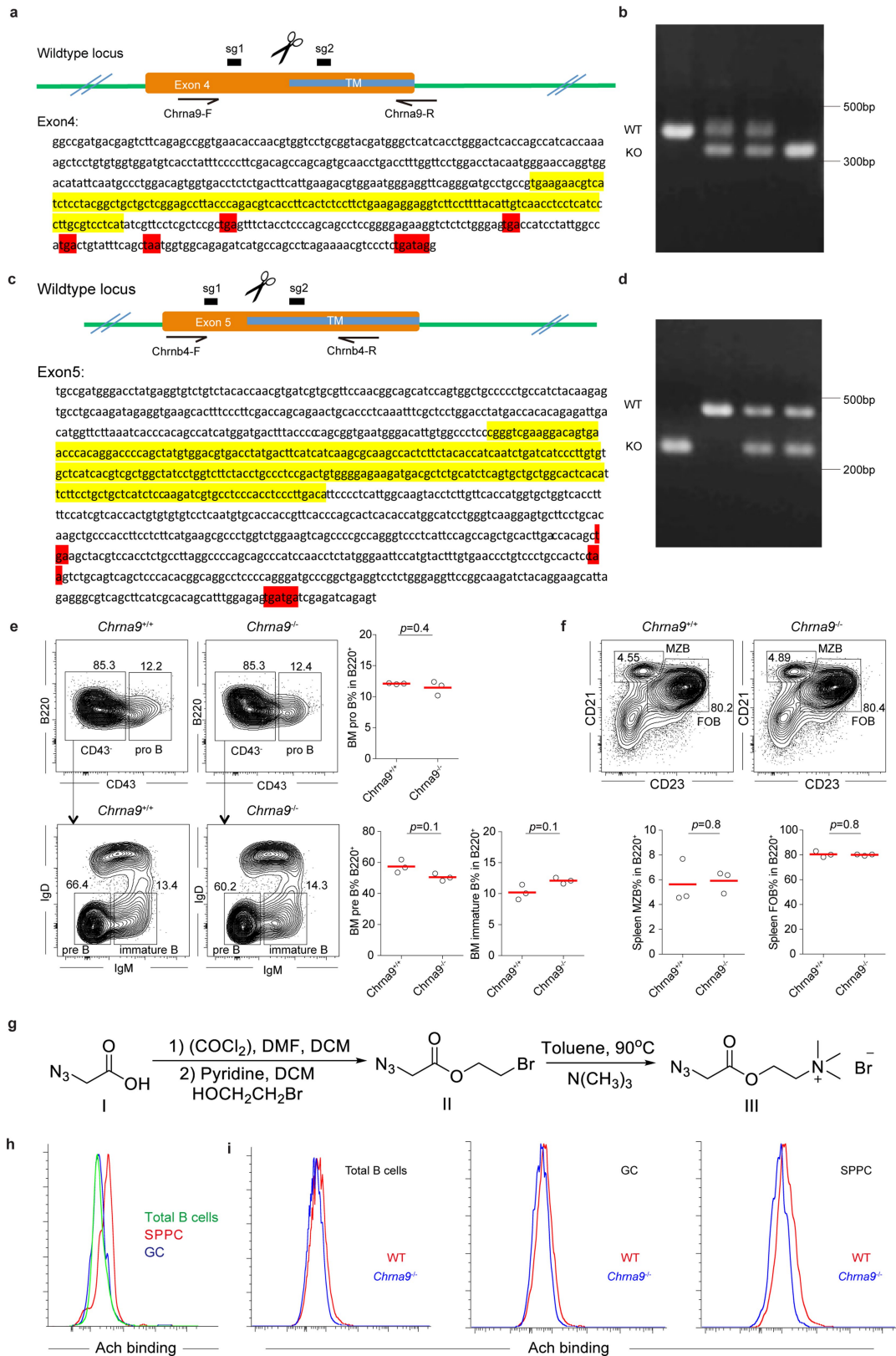


Extended Data Fig. 2 | Effects of acetylcholine on plasma cell formation.

a, b, Percentage SPPC (gated as in Fig. 1d) and GC (gated as in Fig. 1c) 13 days after NP-KLH immunization in mice that were injected subcutaneously with acetylcholine chloride (**a**) or noradrenaline bitartrate monohydrate (**b**) at the indicated doses twice daily from day 8 to day 12 after immunization. Data pooled from three independent experiments; each symbol indicates one mouse, lines denote means. One-way ANOVA. **c, d**, SPPC formation.

c, Representative contour plots showing gated plasma cells in donor CD45.1B

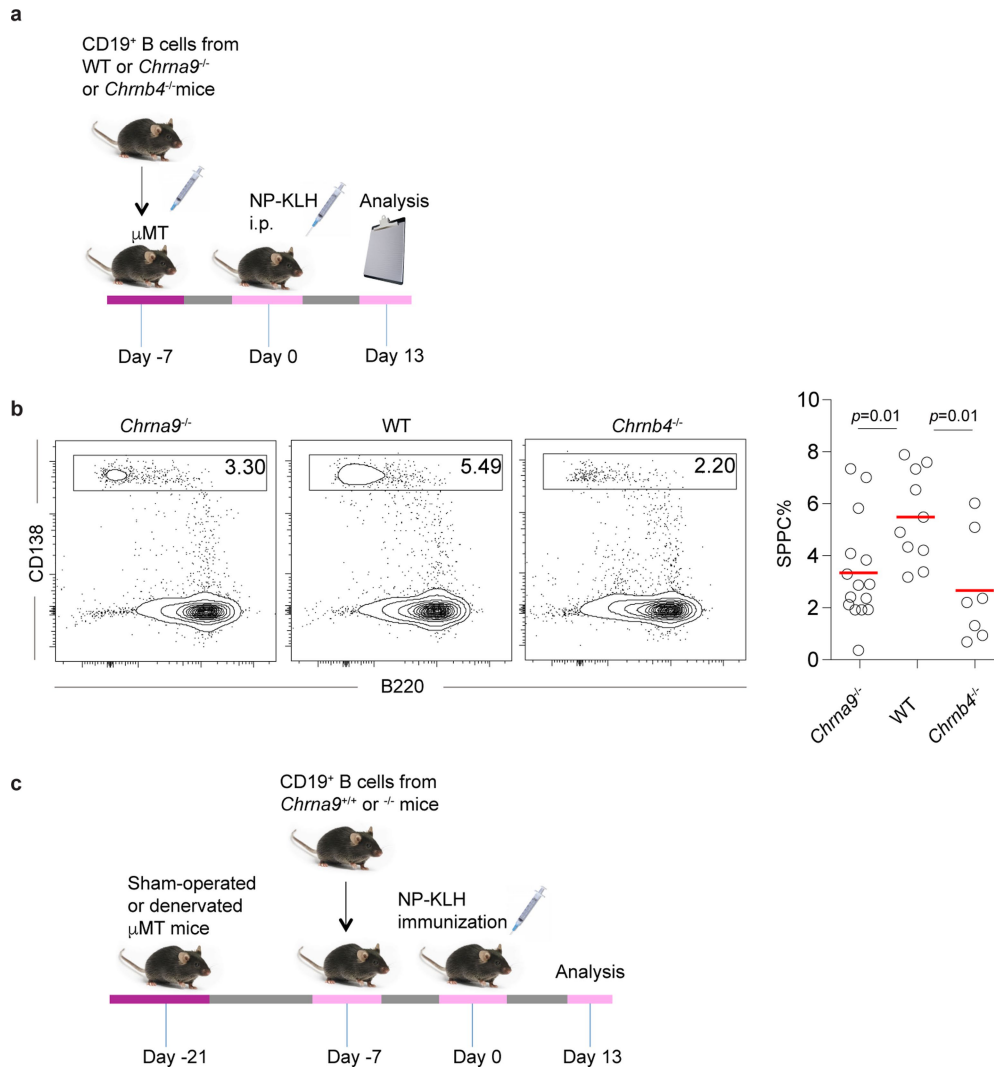
cells that were transduced (GFP⁺) with the AchE-expressing, control, or COMT-expressing vector (top row), donor CD45.1B cells that were not transduced (GFP⁻) (middle row), or recipient B cells of the same recipient host (CD45.2) after NP-KLH immunization (bottom row). **d**, Summary statistics of percentage SPPC in transduced donor B cells (top), non-transduced donor B cells (middle) and recipient B cells (bottom). Data were pooled from three independent experiments. Each symbol indicates one mouse, lines indicate means. All transduction efficiencies were 10–15%. Two-sided unpaired *t*-test.



Extended Data Fig. 3 | See next page for caption.

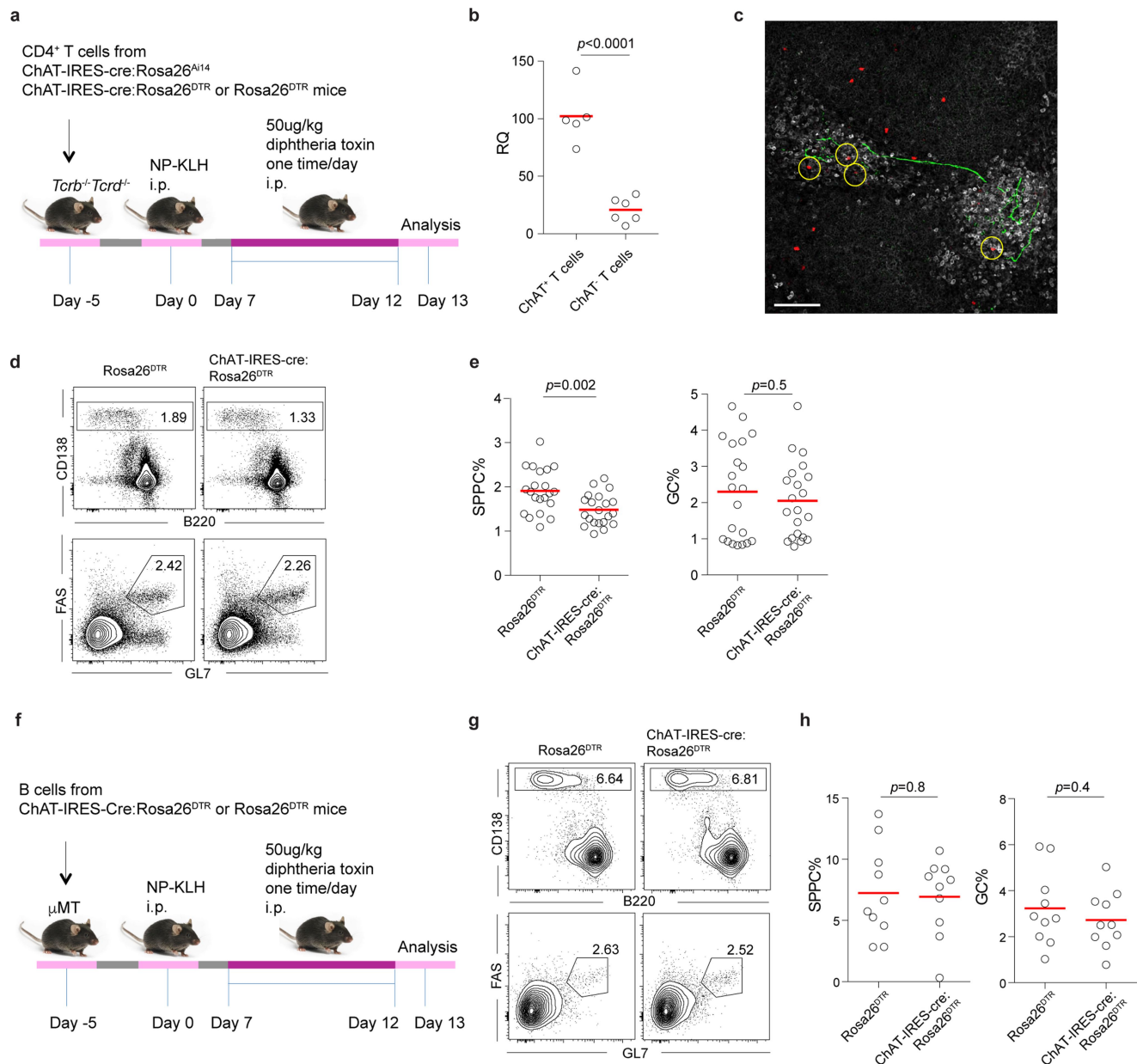
Extended Data Fig. 3 | Generation of mice deficient in AChRs and detection of acetylcholine binding by acetylcholine analogue. a–d, Schematic of *Chrna9* exon 4 (**a**) and *Chrn4* exon 5 (**c**), containing transmembrane domain-coding sequences, with relative positions of the targeting guide RNAs (sg1, sg2) and the genotyping primers. DNA sequences of relevant *Chrna9* (**a**) or *Chrn4* (**c**) gene segments, with the yellow highlight indicating deleted sequences and the red highlight indicating stop codons resulting from the frameshift. PCR genotyping results of *Chrna9* (**b**) or *Chrn4* (**d**) wildtype (WT) and knockout (KO) alleles. **e, f,** Bone-marrow B cell development (**e**) and splenic marginal zone B cell development (**f**) in *Chrna9*^{+/+} and *Chrna9*^{-/-} mice.

Representative FACS contour plots and summary data in scatter plots from one of two experiments. Each symbol indicates one mouse, lines indicate means. Two-sided unpaired *t*-test. **g,** Reaction scheme for synthesis of the acetylcholine analogue, 2-(2-azidoacetoxy)-N,N,N-trimethylethan-1-aminium bromide (III). See Methods for details. **h, i,** Comparison of acetylcholine-binding capacities of wild-type GC, SPPC and total B cells (**h**) and comparison of acetylcholine-binding capacities of *Chrna9*^{+/+} and *Chrna9*^{-/-} cells of indicated types (**i**), as measured by staining with the acetylcholine analogue, chemical III. One of three experiments with similar results is shown.



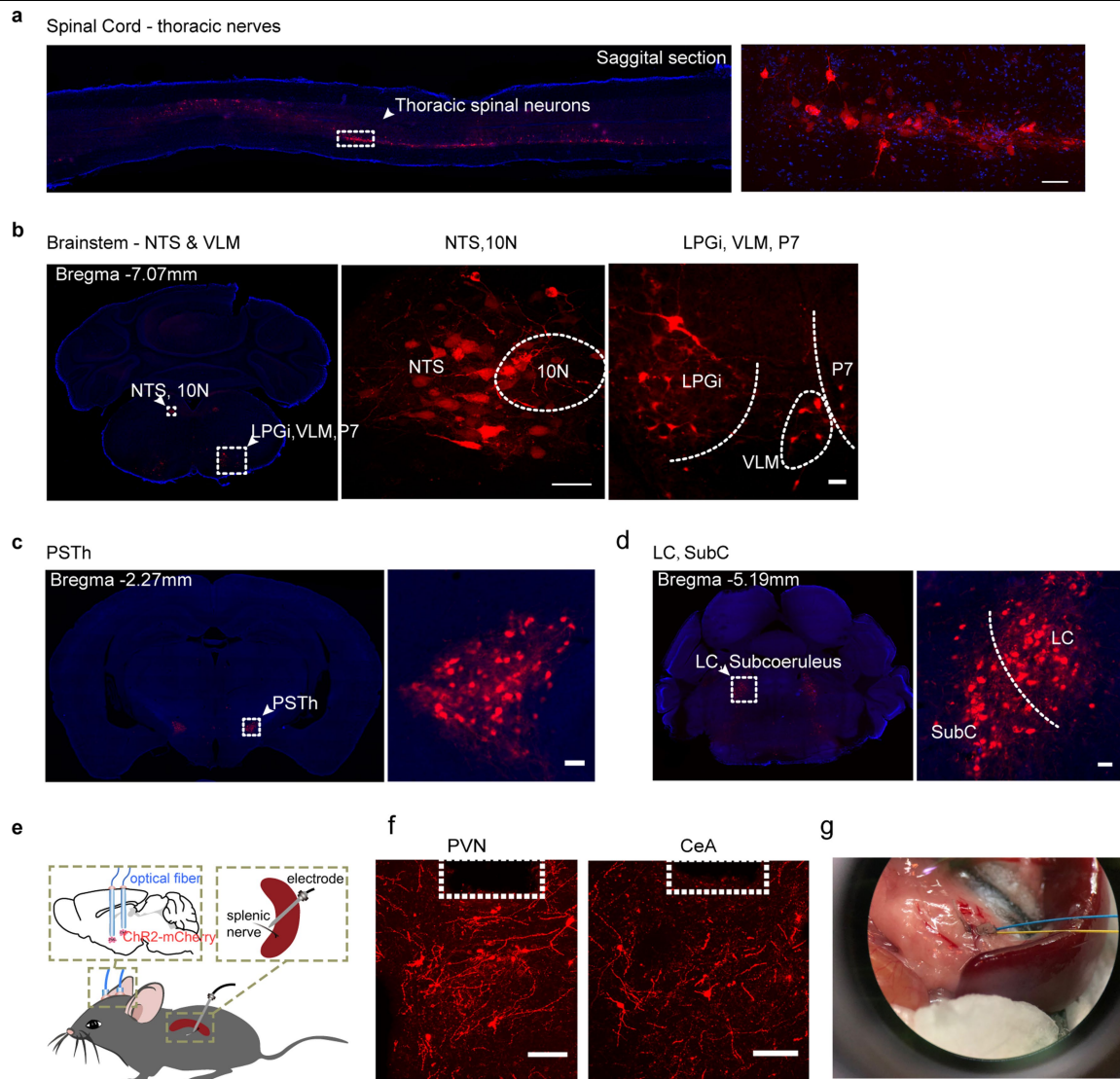
Extended Data Fig. 4 | Requirement for acetylcholine receptors for optimal SPPC formation. **a**, μMT mice were reconstituted by intravenous transfer of wild-type, *Chrna9*^{-/-} or *Chrb4*^{-/-} mature B cells one week before NP-KLH immunization, and percentage SPPC was measured 13 days after immunization. **b**, Representative contour plots and summary statistics of percentage SPPC in reconstituted μMT mice following immunization, as

described in **a**. Data pooled from three independent experiments, two of which had wild-type, *Chrna9*^{-/-} and *Chrb4*^{-/-} donors and one of which had no *Chrb4*^{-/-} group. Points denote individual mice, lines denote means. Two-sided unpaired *t*-test. **c**, The protocol to interrogate the dependence of *Chrna9* effects on splenic nerve, as used in Fig. 2d, e.



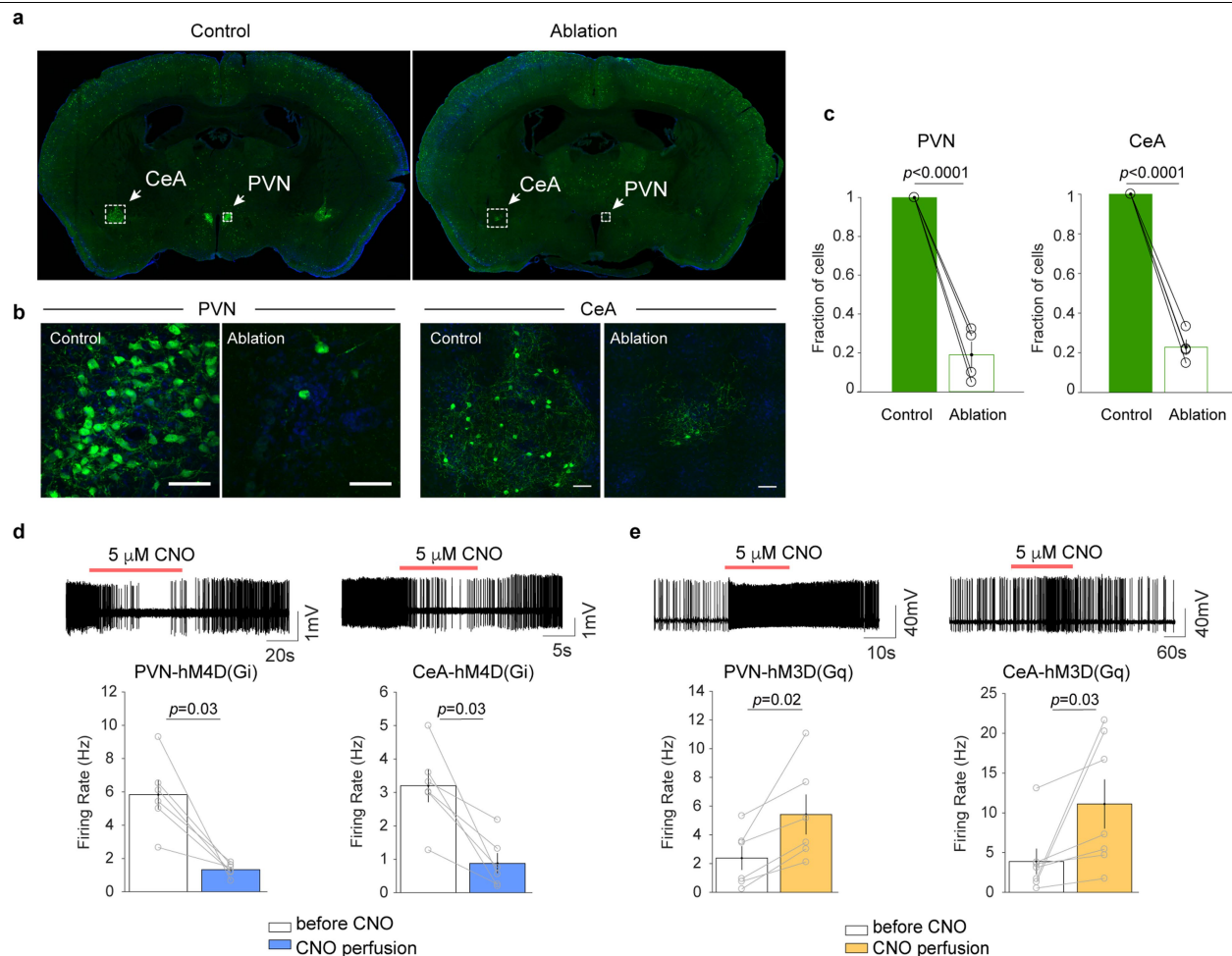
Extended Data Fig. 5 | Involvement of T cells but not B cells as a potential source of acetylcholine for promoting SPPC formation. **a**, The protocol for examining the contribution and function of ChAT⁺ CD4⁺ T cells. **b**, Relative levels of *Chat* mRNA in ChAT⁺ T cells (CD4⁺CD44⁺TdTomato⁺) and ChAT⁺ T cells (CD4⁺CD44⁺TdTomato⁻) sorted from *Tcrb*^{-/-}*Tcrd*^{-/-} knockout mice that were reconstituted with intravenously infused ChAT-IRES-Cre × Rosa26-Δi14 mature T cells and immunized with NP-KLH at day 8. Data were pooled from two independent experiments, with each symbol indicating one independent sort and lines indicating means. **c**, Colocalization (yellow circles) of some ChAT⁺ T cells (red), splenic nerve fibres (TH⁺, green) and aggregates of SPPCs (Igk^{high}, white) on splenic tissue sections taken 10 days after NP-KLH immunization.

Representative of two experiments. Scale bar, 200 μm. **d**, **e**, Representative contour plots (**d**) and summary data (**e**) of percentages of SPPC and GC in DT-treated mice of the indicated genotypes by the protocol shown in **a**. Data pooled from four independent experiments with each symbol indicating one mouse and the lines denoting the means. **f**, The protocol for examining the contribution and function of ChAT⁺ B cells. **g**, **h**, Representative contour plots (**g**) and summary data (**h**) of percentages of SPPC and GC in DT-treated mice of the indicated genotypes by the protocol shown in **f**. Each symbol indicates one mouse from two independent experiments, and lines denote means. Two-sided unpaired *t*-test (**b**, **e**, **h**).



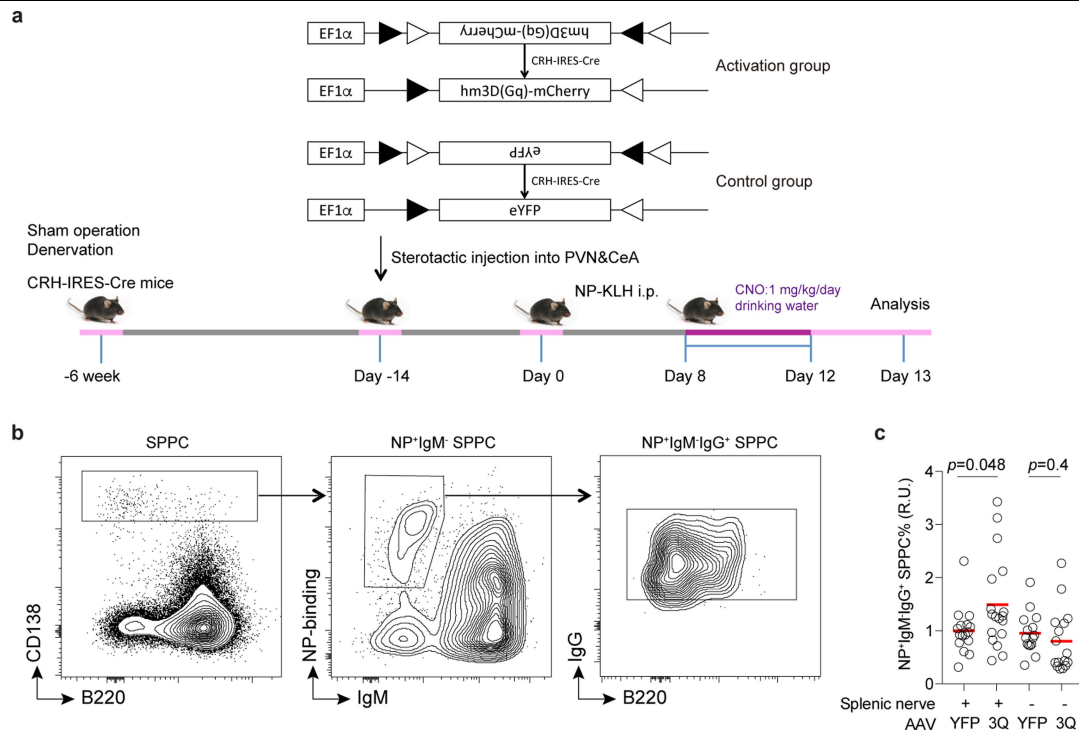
Extended Data Fig. 6 | Retrograde PRV tracing from the spleen and setup for splenic nerve recording. **a–d**, Selected CNS regions that were labelled by PRV. Scale bars, 50 μ m. PSTh, paraventricular nucleus; LC, locus coeruleus; SubC, subcoeruleus nucleus; NTS, nucleus tractus solitarius; VLM, ventrolateral medulla; LPGi, lateral paraventricular nucleus. **e–g**, Setup for optogenetic CRH neuron stimulation and splenic nerve recording in

CRH-IRES-Cre mice. **e**, Procedural diagram for recording splenic nerve activities during optogenetic activation of CRH neurons in the CeA and PVN. **f**, Optical fibre implantation to the PVN and CeA regions in the brain. Red, ChR2⁺ neurons. Scale bars, 50 μ m. **g**, Electrode implantation around the splenic nerve. Data are representative of two independent experiments.



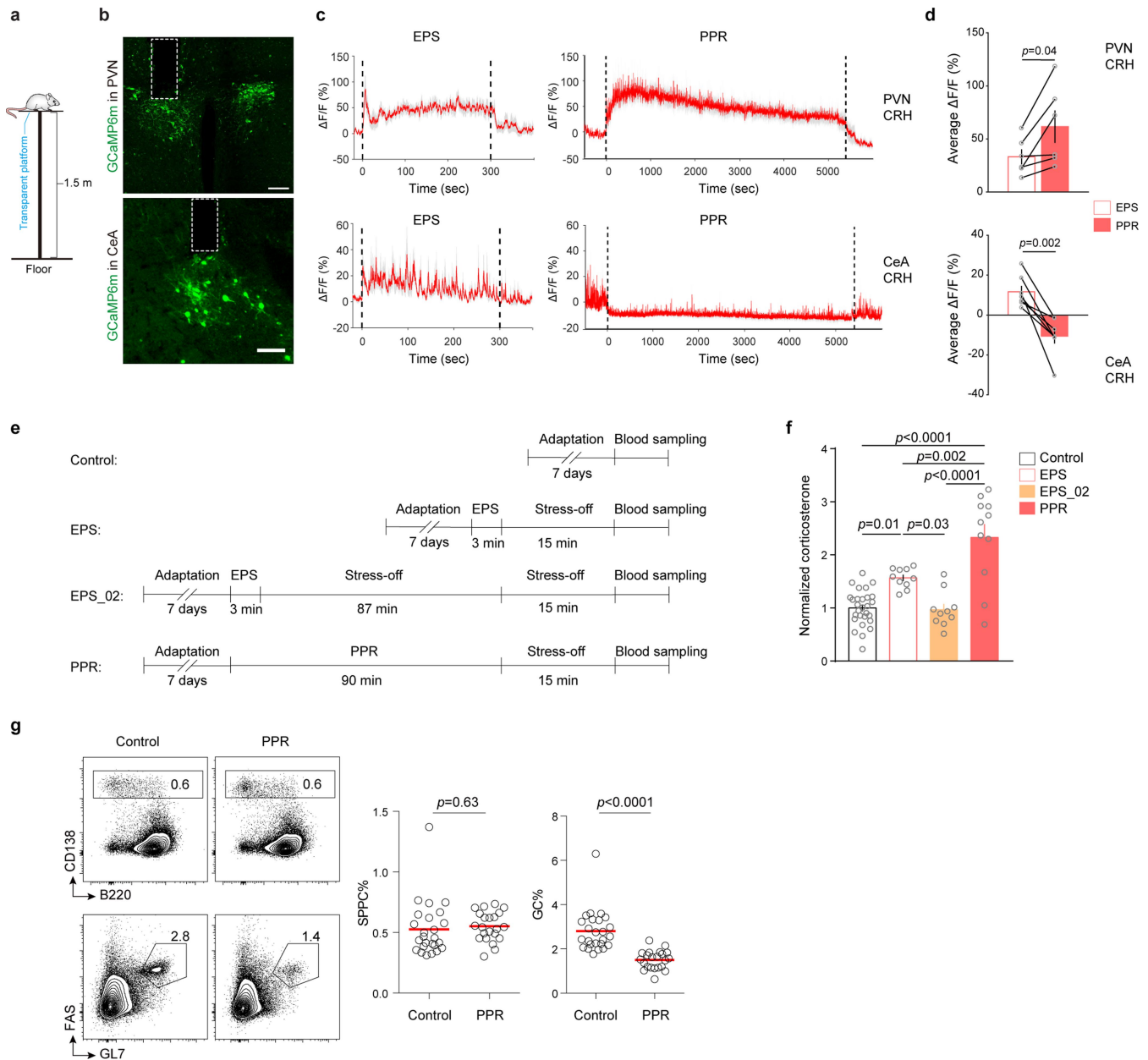
Extended Data Fig. 7 | Efficiency of ablation of CRH neurons and functional verification of DREADD chemogenetics. a–c, The efficiency of CRH neuron ablation. **a,** Representative coronal sections containing CeA and PVN regions from CRH-IRES-Cre:Rosa26^{Al3} mice sham-injected (Control) or injected with AAV-caspase 3 (Ablation) into the CeA and PVN (boxes). **b,** Magnified views of the boxed areas in **a**. Scale bars, 50 μ m. **c,** The relative abundance of Ai3⁺ cells in the PVN and CeA in control (green) or CRH-ablated mice (open bars), with cell numbers in the control group set as 1. Data from four pairs of mice in four independent experiments. Two-sided unpaired *t*-test. **d, e,** Validation of

hM4D(Gi)-mediated inhibition (**d**) and hM3D(Gq)-mediated activation (**e**) by CNO in brain slices. Top, representative recordings from single CRH-IRES-Cre neurons that expressed hM4D(Gi)-mCherry for inhibition (**d**) or hM3D(Gq)-mCherry for activation (**e**) (AAV details in Fig. 3b). Bottom, firing rates of single hM4D(Gi)- or hM3D(Gq)-expressing CRH neurons from indicated regions, before and after perfusion of the brain slices with CNO (5 μ M). Each line with connected symbols indicates one neuron before and after CNO; data collected from three mice in two experiments. Two-sided paired *t*-test.



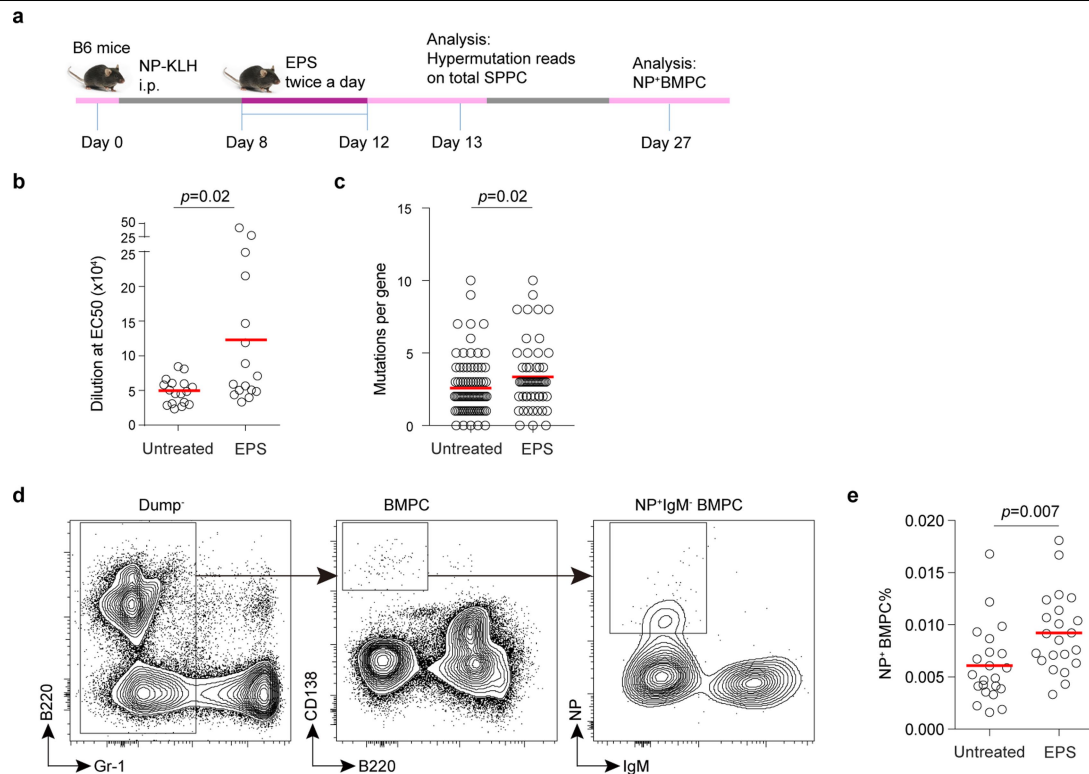
Extended Data Fig. 8 | CRH neuron-mediated promotion of plasma cell formation requires the splenic nerve. **a**, A schematic diagram of the experimental setup. **b**, The gating strategy for NP-specific IgG⁺ SPPCs, from a parental gate of B-lineage cells as in Fig. 1b. **c**, Batch-normalized NP-specific percentage SPPC in relative units (R.U.). For each experiment, the average SPPC% in the Sham-eYFP group is set as unit 1, against which percentage SPPC

in individual mice of the four groups in the same experiment was normalized to obtain the R.U. for each animal. R.U. data are pooled from three independent experiments. Each symbol denotes one mouse, and lines indicate means. Two-sided unpaired *t*-test. YFP, AAV-DIO-eYFP group, 3Q, AAV-DIO-hm3D(Gq) group.



Extended Data Fig. 9 | Distinct stress and immunomodulatory effects of EPS and PPR. **a**, Diagram of the EPS setup. **b**, GCaMP6m expression in the PVN (top) or CeA (bottom) of CRH-IRES-Cre mice as the result of stereotactic AAV injection and positions of implanted optic fibres, outlined in boxes, for photometric measurement of calcium signals in the respective regions. Scale bar, 100 μ m. **c**, Representative traces of integrated calcium signals, presented as normalized changes in the GCaMP6m fluorescence intensity (mean in red, s.e.m. in grey), from CRH neurons in the PVN (top) or CeA (bottom) of CRH-IRES-Cre mice undergoing EPS and then PPR, 3 days apart. Dashed lines mark the beginning and end of behavioural regimens. **d**, Average GCaMP6m

signals during EPS and PPR sessions. Each line denotes one mouse; column heights and error bars show mean \pm s.e.m. of six mice. Two-sided paired *t*-tests. **e**, The schedule of collecting blood samples from mice subjected to EPS (two collection points as indicated) or PPR for corticosterone measurement. **f**, Normalized serum corticosterone levels. One-way ANOVA with Bonferroni's correction. **g**, Representative contour plots (left) and summary statistics of percentage SPCC and GC (right), 13 days after immunization with NP-KLH in mice that were untreated or subjected to PPR twice daily between day 8 and day 12. Data pooled from four independent experiments, with each symbol indicating one mouse and lines indicating means. Two-sided unpaired *t*-test.



Extended Data Fig. 10 | A replication study of the effects of EPS on humoral immunity over 4 weeks. a, Schematic of the experiment setup. **b**, NP-specific IgG in control and EPS-treated mice, quantified as the dilution factor at EC₅₀, calculated for individual mice as in Fig. 5. Each symbol denotes one mouse (17 untreated, 16 EPS), and lines indicate means. **c**, Mutations per V_H186.2 heavy chain in SPPCs on day 13. Each symbol represents one unique V_H186.2 sequence

recovered, and lines indicate means. **d, e**, Gating strategy for (**d**) and frequencies of (**e**) NP-binding B220⁺CD138⁺IgM⁺ plasma cells in the bone marrow. Each symbol represents one mouse, and lines indicate means. Data pooled from three experiments. Two-sided unpaired Mann-Whitney test (**b, c**) or *t*-test (**e**).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Flowcytometry data were collected using BD FACSDIVA V8.0.1.
Imaging data were collected using FV10-ASW V3.1.

Data analysis

Flowcytometry data were processed and analyzed using FlowJo V10.
Imaging data were processed and analyzed using Imaris V7.6.5.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data generated here are included within the paper (and its Supplementary Information files) or available from the corresponding authors upon reasonable request. Source data for Figs. 1–5 and Extended Data Figs. 1–10 are provided with the paper.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size for each experiment is self-evident in the figure or additionally indicated in the legend. The sample size was chosen empirically, based on common experience in the relevant fields, to provide a sufficient level of statistical power for detecting indicated biological effects. No statistical methods were used to pre-determine the sample size.
Data exclusions	No data are excluded.
Replication	All experiments were done at least 3 times, unless explicitly indicated otherwise. Even when data from multiple independent experiments were pooled to conduct statistical group-to-group comparisons, the group-to-group comparison in every experiment actually follows the same trend without exception.
Randomization	For sham operation vs splenic denervation, for stereotactic injection of different viral constructs, for different behavior regimens, co-housed littermates are randomly assigned to groups in comparison.
Blinding	No blinding was involved, as there was no subjective measurement in our experiments.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National

Research sample	Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.
Did the study involve field work?	<input type="checkbox"/> Yes <input type="checkbox"/> No

Field work, collection and transport

Field conditions	Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).
Location	State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).
Access and import/export	Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).
Disturbance	Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Every antibody used is already listed in the Method section, with supplier, clone# when applicable, and catalog# provided. For detail, APC-Cy7 anti-CD19 (1D3), AF700 or BV421 anti-CD4 (GK1.5), PE anti-CD8 (53-6.7), Af700 anti-CD3 (17A2), Af700 anti-CD11c (HL3), biotin-CD43 (S7), PE anti-CD19 (6D5), AF700 or APC-cy7 anti-B220 (RA3-6B2), PE-Cy7 anti-CD93 (A4.1), FITC anti-CD23 (B3B4), APC anti-CD21 (7E9), PE-Cy7 anti-Fas (Jo2), FITC or APC anti-GL7 (GL-7), BV510 or PE or PE-Cy7 anti-CD138 (281-2), EF450 anti-IgM (EB121-15F9), APC and FITC anti-IgM (II/41), BV421 anti-IgM (eB121-15F9), percpCy5.5 anti-IgD (11-26c.2a), EF450 anti-IgD (11-26C), FITC anti-IgG (poly4053), Pacific Blue anti-CD45.1 (A20), APC-Cy7 anti-CD45.2 (104), V450 anti-Gr1 (RB6-8C5) and APC streptavidin (Biolegend), EF450 Streptavidin (eBioscience) and NP-PE (Biosearch Technologies).
Validation	All antibodies are validated by the manufacturers.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	<i>State the source of each cell line used.</i>
Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See ICLAC register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

Palaeontology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	C57BL/6J (Jax 664), CD45.1 (Jax 2014), MT (Jax 2288), Tcrb-/-Tcrd-/- (Jax 2122), ChAT-IRES-Cre (Jax 31661), CRH-IRES-Cre (Jax 12704), Rosa26-iDTR (Jax 7900), Rosa26-Ai3 (Jax 7903) and Rosa26-Ai14 (7914) mice were originally from the Jackson Laboratory and maintained on a B6 background. Relevant mice were interbred to obtain CRH-IRES-Cre:Rosa-Ai3, ChAT-IRES-Cre:Rosa26-iDTR and ChAT-IRES-Cre:Ai14 mice. All mice were housed as groups of 4 to 6 individuals per cage and maintained on a 12-hour light-dark cycle at 22–25°C under specific-pathogen free conditions. All animal experiments were approved by the Institutional Animal Care and Use Committee in accordance of governmental and Tsinghua guidelines for animal welfare. When relevant and applicable, age- and sex-matched mice were randomly chosen from same cages to be included in experimental and control groups.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	All experiments were approved by the Tsinghua University Institutional Animal Care and Usage Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<i>Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."</i>
Recruitment	<i>Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.</i>
Ethics oversight	<i>Identify the organization(s) that approved the study protocol.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.</i>
-----------------------------	---

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Single-cell suspension of the spleen or bone marrow were incubated in MACS buffer (PBS supplemented with 1% FBS and 5 mM EDTA) containing 20 µg/ml 2.4G2 antibody (BioXcell) for 20 min and then stained with indicated antibodies.

Instrument

LSR II cytometer (BD); cell sorting by Aria IV (BD)

Software

FlowJo V10

Cell population abundance

At least 10000 events were collected for cells in the parental gate.

Gating strategy

FSC-A/ SSC-A and 7AAD staining were used to identify viable cells. Singlet cells were identified using FSC-H/ FSC-W gating. Isotype control was used to distinguish between background and marker-positive events.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	<i>Indicate task or resting state; event-related or block design.</i>
Design specifications	<i>Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.</i>
Behavioral performance measures	<i>State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).</i>

Acquisition

Imaging type(s)	<i>Specify: functional, structural, diffusion, perfusion.</i>
Field strength	<i>Specify in Tesla</i>
Sequence & imaging parameters	<i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i>
Area of acquisition	<i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>
Diffusion MRI	<input type="checkbox"/> Used <input type="checkbox"/> Not used

Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	<input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both
Statistic type for inference (See Eklund et al. 2016)	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

Models & analysis

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity
<input type="checkbox"/>	<input type="checkbox"/> Graph analysis
<input type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>

Composition-dependent thermodynamics of intracellular phase separation

<https://doi.org/10.1038/s41586-020-2256-2>

Received: 10 October 2019

Accepted: 1 April 2020

Published online: 6 May 2020

 Check for updates

Joshua A. Riback^{1,6}, Lian Zhu^{1,6}, Mylene C. Ferrolino², Michele Tolbert², Diana M. Mitrea^{2,5}, David W. Sanders¹, Ming-Tzo Wei¹, Richard W. Kriwacki^{2,3,4} & Clifford P. Brangwynne^{1,3,4}✉

Intracellular bodies such as nucleoli, Cajal bodies and various signalling assemblies represent membraneless organelles, or condensates, that form via liquid–liquid phase separation (LLPS)^{1,2}. Biomolecular interactions—particularly homotypic interactions mediated by self-associating intrinsically disordered protein regions—are thought to underlie the thermodynamic driving forces for LLPS, forming condensates that can facilitate the assembly and processing of biochemically active complexes, such as ribosomal subunits within the nucleolus. Simplified model systems^{3–6} have led to the concept that a single fixed saturation concentration is a defining feature of endogenous LLPS^{7–9}, and has been suggested as a mechanism for intracellular concentration buffering^{2,7,8,10}. However, the assumption of a fixed saturation concentration remains largely untested within living cells, in which the richly multicomponent nature of condensates could complicate this simple picture. Here we show that heterotypic multicomponent interactions dominate endogenous LLPS, and give rise to nucleoli and other condensates that do not exhibit a fixed saturation concentration. As the concentration of individual components is varied, their partition coefficients change in a manner that can be used to determine the thermodynamic free energies that underlie LLPS. We find that heterotypic interactions among protein and RNA components stabilize various archetypal intracellular condensates—including the nucleolus, Cajal bodies, stress granules and P-bodies—implying that the composition of condensates is finely tuned by the thermodynamics of the underlying biomolecular interaction network. In the context of RNA-processing condensates such as the nucleolus, this manifests in the selective exclusion of fully assembled ribonucleoprotein complexes, providing a thermodynamic basis for vectorial ribosomal RNA flux out of the nucleolus. This methodology is conceptually straightforward and readily implemented, and can be broadly used to extract thermodynamic parameters from microscopy images. These approaches pave the way for a deeper understanding of the thermodynamics of multicomponent intracellular phase behaviour and its interplay with the nonequilibrium activity that is characteristic of endogenous condensates.

To determine the thermodynamics of LLPS for intracellular condensates, we first focused on the liquid granular component of nucleoli within HeLa cells—in particular on the protein nucleophosmin (NPM1), which is known to be a key driver of nucleolar phase separation^{11,12}. Under typical endogenous expression levels, we estimate the concentration of NPM1 in the nucleoplasm (C^{dil}) to be approximately 4 μM ; from simple binary phase separation models (regular solution theory)¹³ (Supplementary Note 1), this apparent saturation concentration, C_{sat} , is expected to be fixed even under varied protein expression levels (Fig. 1c). Consistent with previous studies¹¹, the overexpression of NPM1 resulted in larger nucleoli, underscoring the importance of NPM1 in nucleolar assembly (Fig. 1a). However, with these increased levels of

NPM1, the nucleoplasmic concentration did not remain fixed at a single C_{sat} , but instead increased by roughly tenfold (Fig. 1b, Supplementary Note 2). Notably, the concentration of NPM1 within the dense-phase nucleolus, C^{den} , also increased, but the ratio of the dense-phase to dilute-phase concentrations, known as the partition coefficient $K = \frac{C^{\text{den}}}{C^{\text{dil}}}$, decreased considerably (Extended Data Fig. 1).

To elucidate the underlying biophysics of this non-fixed C_{sat} within living cells, we examined the phase separation of model biomimetic condensates that are not native within the cell. Using the optoDroplet system⁴ developed for controlling intracellular phase separation, we fused the blue-light-dependent higher-order oligomerizing protein Cry2 to the intrinsically disordered region of DDX4, which drives the

¹Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA. ²Department of Structural Biology, St Jude Children's Research Hospital, Memphis, TN, USA.

³Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. ⁴Howard Hughes Medical Institute, Princeton University, Princeton, NJ, USA. ⁵Present address:

Dewpoint Therapeutics, Boston, MA, USA. ⁶These authors contributed equally: Joshua A. Riback, Lian Zhu. ✉e-mail: richard.kriwacki@stjude.org; brangwy@princeton.edu

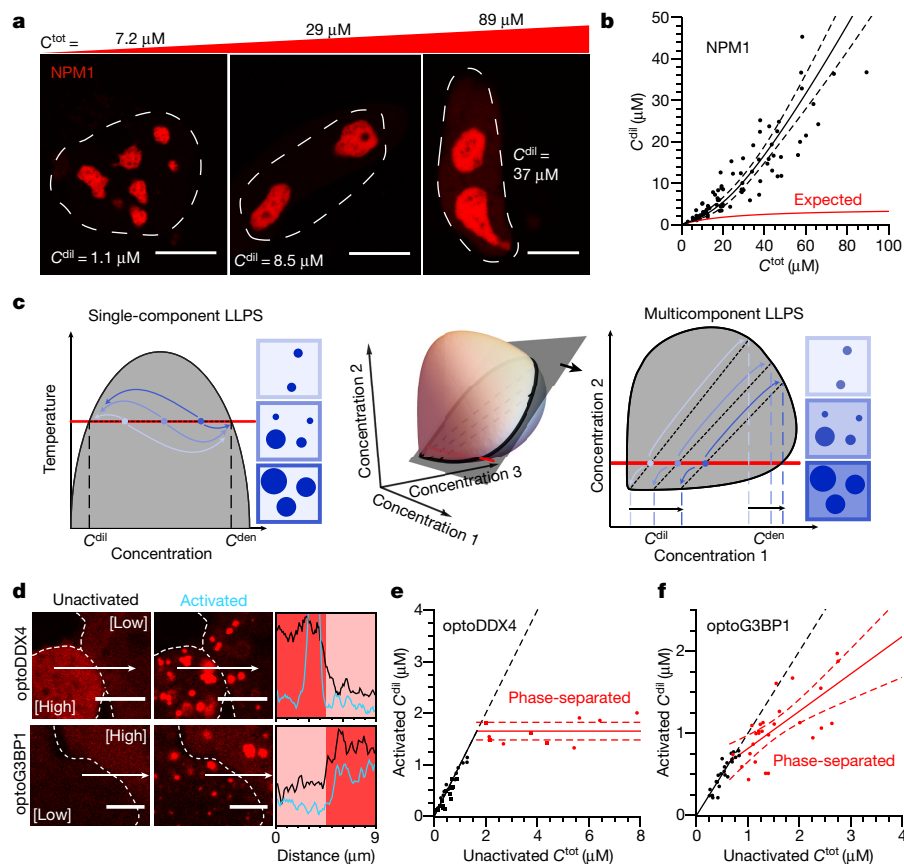


Fig. 1 | Multicomponent LLPS results in non-fixed C_{sat} and the emergence of a concentration-dependent phase stability. **a**, Example images of cells (from $n = 79$ cells) expressing NPM1-mCherry. The total nuclear concentration (C^{tot}) and nucleoplasmic concentration (C^{dil}) of NPM1-mCherry is shown at top of the image and within the image, respectively. The white dashed lines denote the nuclear boundary as defined by NPM1. Scale bars, 10 μm . **b**, The concentration of NPM1-mCherry in the nucleoplasm (C^{dil}) with respect to the total NPM1-mCherry concentration in the nucleus (C^{tot}). The expected trend for a single C_{sat} is shown in red. **c**, Graphical representation of phase diagrams for both single-component (left) and multicomponent (right) LLPS showing fixed and non-fixed C^{dil} (or C_{sat}), respectively. Component concentration changes along the red line; within the grey-shaded region, molecules separate into two phases in which concentrations (curved arrows) are defined by the dashed tie lines. For a multicomponent system, the two-dimensional phase diagram is a slice of a

higher dimensional one, resulting in skewed tie lines and non-fixed C_{sat} . **d**, Example images of cells expressing optoDroplet constructs with optoDDX4 (top, from $n = 19$ cells) or optoG3BP1 (bottom, from $n = 49$ cells), before (left) and after (right) full activation. The line scans shown on the far right correspond to intensity traces before (black) and after (blue) activation. **e**, **f**, Quantification of optoDroplet constructs with optoDDX4 (**e**) and with optoG3BP1 (**f**). The circles represent cytoplasmic concentrations and the squares represent nucleoplasmic concentrations. Cells shown as red points exhibit condensates upon activation (none had condensates before activation); dashed lines represent the mean confidence intervals for cells with foci for constant and linear fits in optoDDX4 and optoG3BP1, respectively. OptoG3BP1 experiments are arsenite-stressed cells in which G3BP1A and G3BP1B are knocked out; optoDDX4 data are reproduced from ref. ¹⁴. Scale bars, 5 μm .

phase separation of exogenous condensates through predominately homotypic interactions^{3,4,10}. Consistent with previous work¹⁴, at total cellular concentrations greater than about 1.7 μM , light activates droplet formation and the nucleoplasmic and cytoplasmic C^{dil} remains at a fixed value, suggesting a fixed C_{sat} of approximately 1.7 μM (Fig. 1d, e). We next asked whether a fixed C_{sat} would be observed upon light induction of stress granules (multicomponent, stress-inducible condensates that assemble through heterotypic protein-mRNA interactions¹⁵). We replaced the oligomerization domain of G3BP1—a critical stress granule protein—with Cry2, and expressed this construct in G3BP1/G3BP2 knockout cells under arsenite stress. At total cytoplasmic concentrations greater than about 0.7 μM , light triggered droplet formation; however, unlike in the case of synthetic DDX4, the C^{dil} was not fixed but instead increased with increasing total concentrations (Fig. 1d, f), similar to the behaviour of NPM1 (Fig. 1a, b). These results are not restricted to light-induced oligomerization of G3BP1 using the optogenetic system, as increasing expression of G3BP1 in a G3BP1/G3BP2 knockout cell line results in a similar increase in the C^{dil} (Extended Data Fig. 2a).

These data suggest that multicomponent condensates are not governed by a fixed C_{sat} , as would be expected for a single-biomolecule-component (that is, binary solution when including the solvent) (Supplementary Note 1) phase boundary at fixed temperature. Instead, endogenous condensates may be governed by the more richly textured thermodynamics that dictate higher-dimensional phase diagrams (Fig. 1c), consistent with theoretical and experimental findings on model multicomponent systems^{13,16–22}. To investigate this concentration-dependent thermodynamics, we quantify the effect of increasing the concentration of a biomolecule in vivo or in vitro, which shifts the stoichiometry to bias towards more homotypic interactions (Fig. 2a, Extended Data Fig. 3, Supplementary Note 3). This changes the partition coefficient, enabling us to quantify changes in the generalized standard free energy of transfer, here denoted as ΔG^{tr} , for any component from the dilute to the dense phase (Fig. 2b); thermodynamic considerations yield the relationship $\Delta G^{\text{tr}} = -RT \ln K$ (Supplementary Note 4). For components that contribute to phase separation (for example, those that act to scaffold the condensate

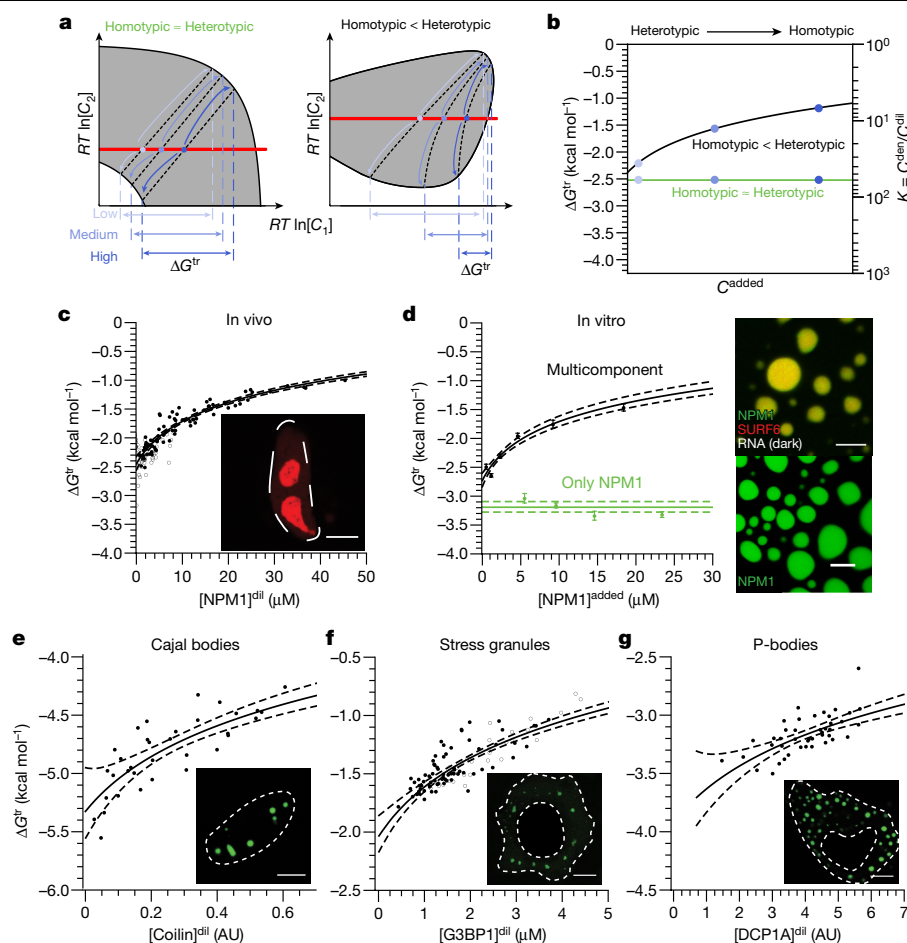


Fig. 2 | Determining the contribution of heterotypic and homotypic interactions that drive condensate formation in vivo and in vitro.

a, Schematic of the connection between the phase diagram and the transfer free energy of a component when heterotypic interactions are equal to (left) or stronger than (right) homotypic interactions. C_1 and C_2 represent components 1 and 2. **b**, Accompanying schematic to **a**, detailing the qualitative change in the transfer free energy of component 1 with an increase in its expression for the two cases in **a**. **c**, Thermodynamic dependence of NPM1 (–mCherry filled, –GFP empty) transfer from the nucleoplasm into the nucleolus, as a function of its increased expression (concentration in the nucleoplasm). The inset is an image

from Fig. 1a, to highlight that these data represent a reanalysis of those experiments. **d**, Left, ΔG^{tr} for NPM1 as a function of added NPM1, obtained from in vitro reconstitution experiments. Right, images of NPM1 droplets with 5% PEG (bottom) and of ternary NPM1:SURF6N:rRNA droplets in buffer (top). **e–g**, ΔG^{tr} for coilin–eYFP (**e**), G3BP1 (**f**, –GFP empty, –mCherry filled), and DCP1A–eYFP (**g**) from the dilute phase (that is, nucleoplasm or cytoplasm) to Cajal bodies, arsenite-induced stress granules, and P-bodies (that is, dense phases), respectively. For all proteins here, a higher C^{dil} results from an increase in its expression (Fig. 1b, Extended Data Fig. 2a–c). AU, arbitrary units. Scale bars, 10 μm.

meshwork), their transfer free energy reports on the stability of interactions driving phase separation.

Applying this framework to our results for NPM1 (Fig. 1a, b) reveals that as the concentration of NPM1 is increased, the partition coefficient of NPM1 into the nucleolus decreases (Extended Data Fig. 1b); as such, the transfer free energy ΔG^{tr} for NPM1 between the condensed and the dilute phases becomes less negative, and thus destabilizing (Fig. 2c). This destabilizing effect at higher NPM1 concentrations implies that heterotypic—rather than homotypic (that is, NPM1–NPM1)—interactions dominate endogenous nucleolar assembly. To further test this conclusion, we focused on in vitro reconstitution of the nucleolar granular component. In addition to NPM1, key granular component biomolecules include ribosomal RNA (rRNA), multivalent proteins containing polyarginine motifs (Arg-proteins, such as SURF6) and ribosomal proteins (r-proteins). Using a well-established system for the phase separation of granular component biomolecules in vitro^{11,12,20,23}, we formed either NPM1-only droplets with 5% PEG as a crowder (Fig. 2d, bottom) or multicomponent droplets containing NPM1, the N terminus of SURF6 (SURF6N) and rRNA (Fig. 2d, top). As expected for single-biomolecule-component phase separation, as more NPM1 was

added to the NPM1-only droplets, the transfer free energy remained roughly constant (Fig. 2d, green). By contrast, for multicomponent droplets, the transfer free energy became substantially less negative (that is, destabilizing) as more NPM1 was added, as was observed in living cells (Fig. 2d, black).

Notably, similar behaviour in cells was observed with numerous different intracellular condensates and their associated key scaffolding proteins: coilin in Cajal bodies, G3BP1 in arsenite-triggered stress granules and DCP1A in P-bodies. In each of these cases, increasing protein concentrations yielded larger condensates, surrounded by a higher C^{dil} , and with correspondingly less-negative transfer free energies (Fig. 2e–g, Extended Data Fig. 2); these data are consistent with previous studies that highlight the complex nature of biomolecule recruitment to in vitro- and in vivo-reconstituted condensates^{12,24}. However, our findings contrast with the view that condensates are stabilized by predominantly homotypic interactions, for example those mediated by self-associating intrinsically disordered regions. Instead, the data suggest that heterotypic interactions have a central role in promoting the internal cohesivity that stabilizes LLPS—not only for nucleoli, but also for other intracellular condensates.

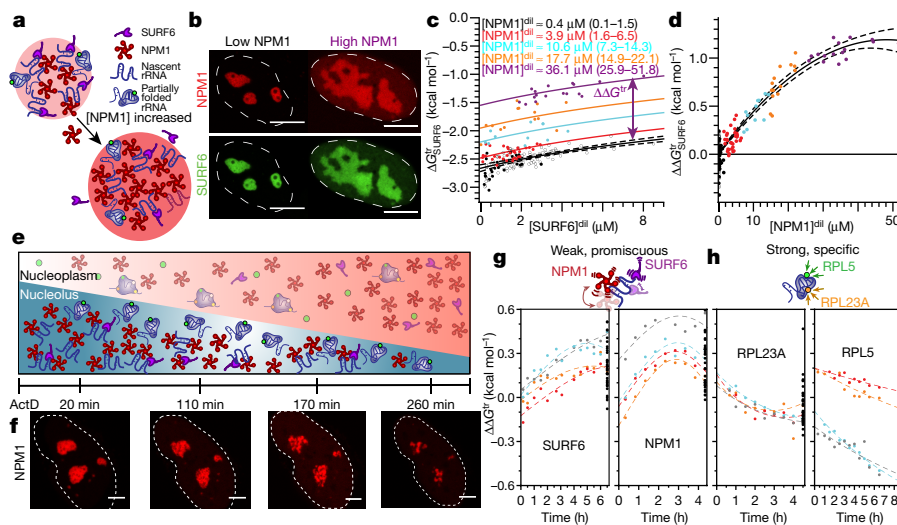


Fig. 3 | Heterotypic interactions between nucleolar proteins and rRNA underlie nucleolar thermodynamics. **a**, Schematic of the proposed mechanism for the dilution of non-NPM1 molecular interactions in the dense phase owing to NPM1 overexpression. Only relevant species are shown for clarity. **b**, Example images of cells (from $n = 102$ cells) expressing NPM1-mCherry (top) and SURF6-GFP (bottom) with low (left) and high (right) expression of NPM1. Scale bar, 10 μm . **c**, Change in the transfer free energy of SURF6 with overexpression of NPM1 plotted against the concentration of SURF6. The colours indicate different concentrations of NPM1 with mean and range values indicated; open circles are cells without additional NPM1 expressed. The method of calculating ΔG^{tr} at a referenced nucleoplasmic

SURF6 concentration is shown via arrows and displaced lines in **c**, **d**. The change in ΔG^{tr} shown as a function of NPM1 concentration; the colour code is the same as in **c**, **e**. **e**, Schematic showing the effect of ActD treatment on nucleoli over time. **f**, Images of cells at the indicated times after ActD treatment (from $n = 4$ NPM1-tagged time series). The corresponding quantification for NPM1 cells is shown in Extended Data Fig. 5. Scale bars, 5 μm . **g**, **h**, ΔG^{tr} of SURF6 and NPM1 (**g**) and RPL23A and RPL5 (**h**) plotted against time after ActD treatment. Each colour represents an individual cell followed over time; black points are cells measured at the indicated time points. The schematics at the top of **g** and **h** highlight the differences in suggested interactions with rRNA.

We next investigated which heterotypic interactions drive phase separation of the nucleolus, by monitoring the transfer free energy of one component while changing the concentration of another (Fig. 3a). In our multicomponent *in vitro* mimic, we found that increasing the concentrations of NPM1 or SURF6 renders the partitioning of SURF6 less energetically favourable (Extended Data Fig. 4); this is again consistent with heterotypic interactions driving SURF6 to nucleoli. In living cells, SURF6 also exhibits behaviour similar to that of NPM1, with a destabilizing increase in the transfer free energy observed with increasing SURF6 concentration (Fig. 3c, black). This *in vivo* destabilization is markedly amplified with increasing NPM1 concentrations (Fig. 3b, c). From these data, we determined the change in ΔG^{tr} of SURF6 as a function of NPM1, by referencing to the energy expected without NPM1 overexpression—that is, $\Delta G^{\text{tr}}_{\text{SURF6}}([\text{NPM1}]^{\text{dil}}) = \Delta G^{\text{tr}}_{\text{SURF6}}([\text{NPM1}]^{\text{dil}}, [\text{SURF6}]^{\text{dil}}) - \Delta G^{\text{tr}}_{\text{SURF6}}([\text{NPM1}]^{\text{dil}} = 0, [\text{SURF6}]^{\text{dil}})$. Plotting $\Delta G^{\text{tr}}_{\text{SURF6}}$ against NPM1 collapses the data onto a single master curve (Fig. 3d, Supplementary Methods), highlighting a tight thermodynamic link between NPM1 and SURF6. This behaviour contrasts with that of r-proteins, which exhibit strong and specific rRNA binding, and a transfer free energy that is statistically insensitive to the concentration of NPM1 (Extended Data Fig. 5).

Both SURF6 and NPM1 have been proposed to interact with rRNA through weak promiscuous binding¹². We therefore suggested that SURF6–NPM1 linkage occurs as a consequence of heterotypic interactions with rRNA, which are diluted upon NPM1 overexpression. To test whether heterotypic interactions with rRNA underlie the thermodynamics of nucleolar assembly, we performed our analysis in cells after treatment with actinomycin D (ActD), which is known to halt the transcription of nascent rRNA without affecting the processing and assembly of pre-existing rRNA^{25,26} (Fig. 3e). As previously reported, the addition of ActD results in a progressive reduction of nucleolus size over the course of 4 hours²⁷ (Fig. 3f, Extended Data Fig. 6). Over time, the ΔG^{tr} of NPM1 and SURF6 increases, indicating weakened interactions relative to cells without ActD treatment. This is consistent

with NPM1 and SURF6 driving heterotypic phase separation through multivalent interactions with nascent, unfolded (or misfolded) rRNA transcripts, which become increasingly scarce under ActD treatment. Conversely, we find that the two r-proteins RPL23A and RPL5 display the opposite behaviour—their transfer free energies decrease as ActD treatment progresses (Fig. 3g, h), reflecting strengthened interactions that are consistent with specific binding to more fully processed rRNA.

These findings shed light on how heterotypic interactions that drive phase separation facilitate sequential rRNA processing in ribosome biogenesis. Specifically, when compared with fully assembled ribosome subunits, relatively nascent rRNA transcripts are available for a greater number of interactions with NPM1, SURF6 and other scaffolding components of the granular component matrix, providing a mechanism to facilitate the vectorial flux of processed subunits out of the nucleolus²⁰ (Fig. 4f). Indeed, binding of nascent transcripts by r-proteins eliminates multivalent binding sites for heterotypic scaffolding proteins, which could serve to effectively expel fully assembled pre-ribosomal particles. We tested this concept using the biomimetic Corelet system—a 24-mer ferritin core in which each ferritin subunit is fused to an optogenetic heterodimerization domain that can be used to tune the effective valency of the particle with light⁶ (Fig. 4a). We fused the optogenetic protein to an N-terminal-truncated construct of NPM1 (NPM1-C; residues 120–294), thereby allowing light-dependent multivalent interactions with the nucleolus. On its own, this construct partitions only weakly into nucleoli, with a ΔG^{tr} of approximately $-0.4 \text{ kcal mol}^{-1}$ (Extended Data Fig. 7). In the absence of bound NPM1-C, the ferritin core is strongly excluded from nucleoli with a ΔG^{tr} of approximately $+1.4 \text{ kcal mol}^{-1}$ (Extended Data Fig. 7); this is consistent with large non-interacting assemblies being excluded from the nucleolus and other condensates^{28–30}. However, increasing the valence of the core by light activation results in an increase in its partitioning into the nucleolus, implying a more favourable (that is, negative) transfer free energy. This effect depends strongly on the valence of the core:

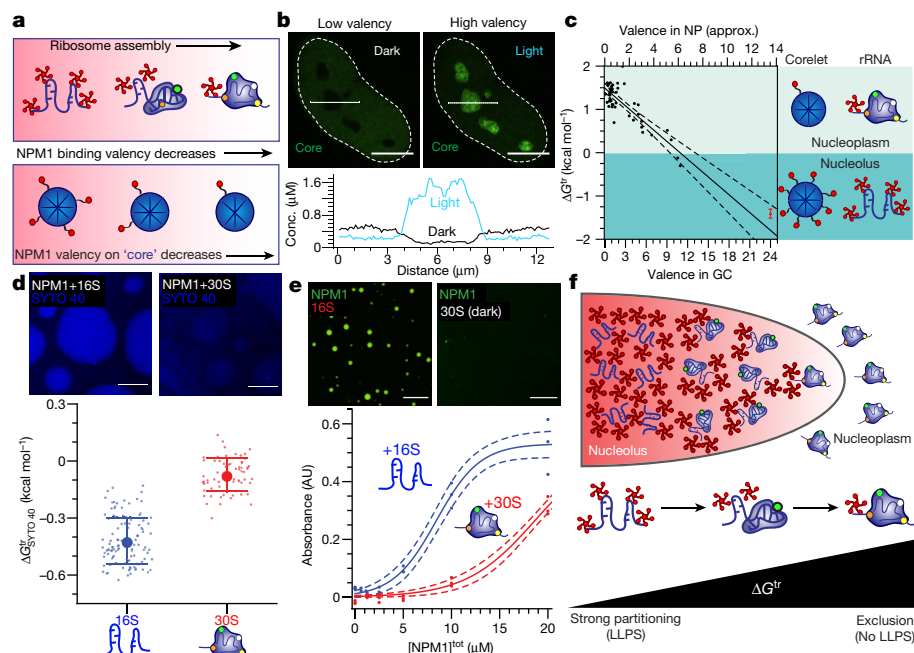


Fig. 4 | Composition-dependent heterotypic LLPS drives specific ribosomal subunit exclusion. **a**, Top, schematic of NPM1 valency as a function of rRNA folding and processing in the nucleolus; bottom, schematic of NPM1 valency on ferritin ‘cores’ using the Corelet optogenetic system. **b**, Images of a cell highlighting the partitioning of the cores before light exposure (low effective valence) (left) and after light exposure (high effective valence) (right), upon which NPM1-C binding sites on the core are saturated in this cell. Quantification is shown below, corresponding to the dashed line shown in the images. **c**, Corresponding quantification of the dependence of the ΔG^{tr} of the core on the valence in the granular component (GC) after light activation. Dotted lines are fits to data. NP, nucleoplasm. **d**, Top, representative images of 16S RNA (left) or the 30S small ribosomal subunit (right) partitioning into pre-formed 10 μM NPM1 droplets (made with 5% PEG-8K); the RNA species

(used at 5 $\mu\text{g ml}^{-1}$) were visualized using 6.5 μM SYTO 40. Bottom, the corresponding transfer free energies of droplet formation. The large circles represent the mean and the error bars represent the standard deviation from $n = 118$ droplets (16S) and $n = 64$ droplets (30S). **e**, Top, microscopy images of 10 μM NPM1 incubated with 16S RNA (left) or the 30S small ribosomal subunit (right). Bottom, turbidity assay of NPM1 incubated at various concentrations with either 16S rRNA or the 30S small ribosomal subunit. The RNA species was added at 50 $\mu\text{g ml}^{-1}$; for validation of protein and RNA components see Extended Data Fig. 8. 16S rRNA was labelled via a morpholino approach as described in the Supplementary Methods. **f**, Proposed mechanism of ribosomal subunit exclusion from the granular component of the nucleolus driven by thermodynamics of nucleolar LLPS.

Corelets of valence less than 10 are excluded from the nucleolus ($\Delta G^{\text{tr}} > 0$), whereas those of valence greater than 10 are enriched ($\Delta G^{\text{tr}} < 0$) within the nucleolus (Fig. 4b, c, Extended Data Fig. 7). This physical picture is supported by in vitro experiments with NPM1 droplets and with ribosomal components of *Escherichia coli*, which reveal that ΔG^{tr} is more strongly negative for 16S rRNA compared with the 30S ribosomal subunit (comprising 16S plus associated r-proteins (Extended Data Fig. 8a–c)) (Fig. 4e). Consistent with these measurements, the in vitro phase separation of NPM1 is substantially weaker in the presence of the 30S subunit compared with 16S rRNA (Fig. 4d); this underscores how non-ribosomal protein bound (that is, smaller and highly solvent-exposed) rRNAs are associated with favourable heterotypic interactions that promote partitioning and phase separation with nucleolar scaffold proteins (Fig. 4d, e). Similarly, in vitro phase separation was substantially weaker in the presence of the full 70S ribosome compared with either 23S rRNA or total (that is, 23S, 16S and 5S) rRNA (Extended Data Fig. 8d). Taken together, these data suggest a mechanism in which phase separation of rRNA with the nucleolar scaffold becomes progressively less energetically favourable as components mature into fully assembled ribosomal subunits, leading to their thermodynamically driven exit from nucleoli.

Our findings lay the groundwork for a quantitative understanding of the interplay between the composition-dependent thermodynamics of condensate assembly and the free-energy landscape of biomolecular complex assembly. In particular, we show that heterotypic biomolecular interactions give rise to high-dimensional phase behaviour that yields C_{sat} values that vary with component concentrations, providing

a mechanism for tuning condensate composition. This enables ‘on demand’ condensate assembly—such that phase separation occurs only in the presence of the substrate—while simultaneously enabling a non-equilibrium steady-state flux of products (substrates), which are driven out of (or in to) the condensate during processing. This is likely to be relevant not only to the nucleolus, but also to many other phase-separated condensates that facilitate the formation of diverse biomolecular complexes, such as the spliceosome. Future work will exploit these intracellular thermodynamic self-assembly principles towards new organelle-engineering applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2256-2>.

- Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382 (2017).
- Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
- Nott, T. J. et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015).
- Shin, Y. et al. Spatiotemporal control of intracellular phase transitions using light-activated optoDroplets. *Cell* **168**, 159–171.e14 (2017).
- Wang, J. et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).
- Bracha, D. et al. Mapping local and global liquid phase behavior in living cells using photo-oligomerizable seeds. *Cell* **175**, 1467–1480.e13 (2018).

7. Holehouse, A. S. & Pappu, R. V. Functional implications of intracellular phase transitions. *Biochemistry* **57**, 2415–2423 (2018).
8. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell* **176**, 419–434 (2019).
9. McSwiggen, D. T., Mir, M., Darzacq, X. & Tjian, R. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes Dev.* **33**, 1619–1634 (2019).
10. Oltsch, F., Klosin, A., Julicher, F., Hyman, A. A. & Zechner, C. Phase separation provides a mechanism to reduce noise in cells. *Science* **367**, 464–468 (2019).
11. Feric, M. et al. Coexisting liquid phases underlie nucleolar subcompartments. *Cell* **165**, 1686–1697 (2016).
12. Mitrea, D. M. et al. Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA. *eLife* **5**, e13571 (2016).
13. Flory, P. J. *Principles of Polymer Chemistry* (Cornell Univ. Press, 1953).
14. Wei, M.-T., Chang, Y.-C., Shimobayashi, S. F., Shin, Y. & Brangwynne, C. P. Nucleated transcriptional condensates amplify gene expression. Preprint at <https://www.biorxiv.org/content/10.1101/737387v2> (2019).
15. Kedersha, N. et al. G3BP–Caprin1–USP10 complexes mediate stress granule condensation and associate with 40S subunits. *J. Cell Biol.* **212**, e201508028 (2016).
16. Choi, J.-M., Dar, F. & Pappu, R. V. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput. Biol.* **15**, e1007028 (2019).
17. Mao, S., Kuldinow, D., Haataja, M. P. & Košmrlj, A. Phase behavior and morphology of multicomponent liquid mixtures. *Soft Matter* **15**, 1297–1311 (2019).
18. Priftis, D. & Tirrell, M. Phase behaviour and complex coacervation of aqueous polypeptide solutions. *Soft Matter* **8**, 9396–9405 (2012).
19. Jacobs, W. M. & Frenkel, D. Phase transitions in biological systems with many components. *Biophys. J.* **112**, 683–691 (2017).
20. Mitrea, D. M. et al. Self-interaction of NPM1 modulates multiple mechanisms of liquid-liquid phase separation. *Nat. Commun.* **9**, 842 (2018).
21. Lin, Y.-H., Brady, J. P., Forman-Kay, J. D. & Chan, H. S. Charge pattern matching as a ‘fuzzy’ mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. *New J. Phys.* **19**, 115003 (2017).
22. Banerjee, P. R., Milin, A. N., Moosa, M. M., Onuchic, P. L. & Deniz, A. A. Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets. *Angew. Chem. Int. Ed.* **56**, 11354–11359 (2017).
23. Ferrolino, M. C., Mitrea, D. M., Michael, J. R. & Kriwacki, R. W. Compositional adaptability in NPM1–SURF6 scaffolding networks enabled by dynamic switching of phase separation mechanisms. *Nat. Commun.* **9**, 5064 (2018).
24. Banani, S. F. et al. Compositional control of phase-separated cellular bodies. *Cell* **166**, 651–663 (2016).
25. Geuskens, M. & Bernhard, W. Cytochimie ultrastructurale du nucléole. 3. Action de l’actinomycine D sur le métabolisme du RNA nucléolaire. *Exp. Cell Res.* **44**, 579–598 (1966).
26. Lazdins, I. B., Delannoy, M. & Sollner-Webb, B. Analysis of nucleolar transcription and processing domains and pre-rRNA movements by in situ hybridization. *Chromosoma* **105**, 481–495 (1997).
27. Burger, K. et al. Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. *J. Biol. Chem.* **285**, 12416–12425 (2010).
28. Wei, M.-T. et al. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **9**, 1118–1125 (2017).
29. Zhu, L. et al. Controlling the material properties and rRNA processing function of the nucleolus using light. *Proc. Natl Acad. Sci. USA* **116**, 17330–17335 (2019).
30. Handwerker, K. E., Cordero, J. A. & Gall, J. G. Cajal bodies, nucleoli, and speckles in the *Xenopus* oocyte nucleus have a low-density, sponge-like structure. *Mol. Biol. Cell* **16**, 202–211 (2005).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Source data for Figs. 1–4 and Extended Data Figs. 1, 2, 3, 4–8 are available with the paper. All other data are available from the corresponding authors upon reasonable request.

Acknowledgements We thank members of the Brangwynne laboratory for discussions and comments on this manuscript. This work was supported by the Howard Hughes Medical Institute, the St Jude Collaborative on Membraneless Organelles, and grants from the National Institutes of Health (NIH) 4D Nucleome Program (U01 DA040601) and the Princeton Center for Complex Materials, a Materials Research Science and Engineering Center supported by the National Science Foundation (NSF) (DMR 1420541). L.Z. was supported by an NSF graduate fellowship (DGE-1656466). R.W.K. acknowledges support from the NIH (R01 GM115634, R35

GM131891 and P30 CA021765 (to St Jude Children's Research Hospital)) and ALSAC. M.T. acknowledges support from the NIH (F32 GM131524). Some images were acquired at the St Jude Cell & Tissue Imaging Center, which is supported by St Jude Children's Research Hospital and the National Cancer Institute (P30 CA021765); we thank V. Frohlich and J. Peters for technical assistance.

Author contributions J.A.R., L.Z., D.M.M., R.W.K. and C.P.B. designed the research; in vivo studies were performed and analysed by J.A.R., L.Z., D.W.S. and M.W.; in vitro studies were performed and analysed by M.C.F., M.T. and D.M.M.; J.A.R., L.Z. and C.P.B. wrote the manuscript, which was reviewed and edited by all authors.

Competing interests R.W.K. is a consultant for, and D.M.M. has recently become employed by, Dewpoint Therapeutics. The remaining authors declare no competing interests.

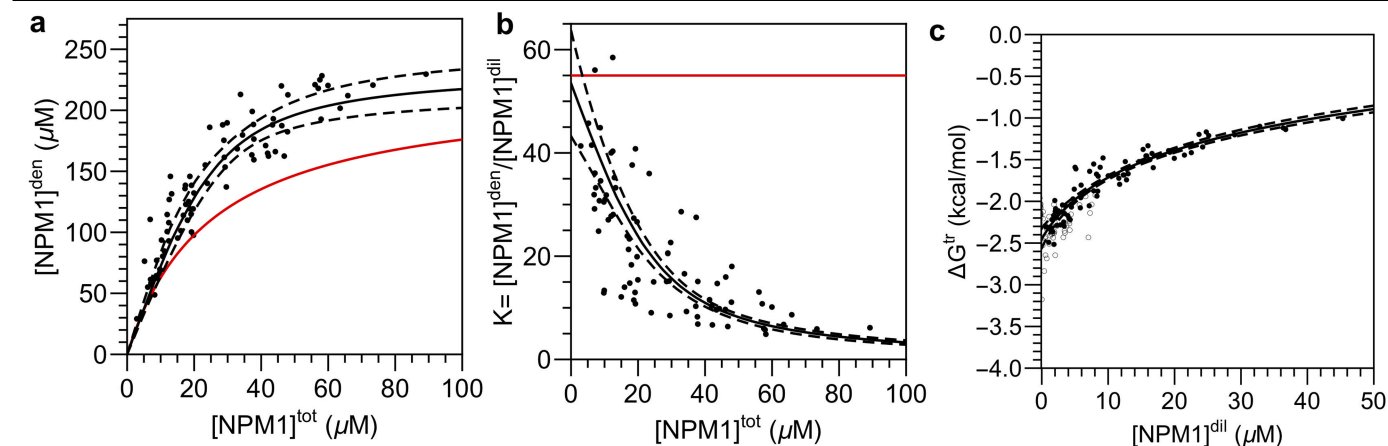
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2256-2>.

Correspondence and requests for materials should be addressed to R.W.K. or C.P.B.

Peer review information *Nature* thanks Rohit Pappu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

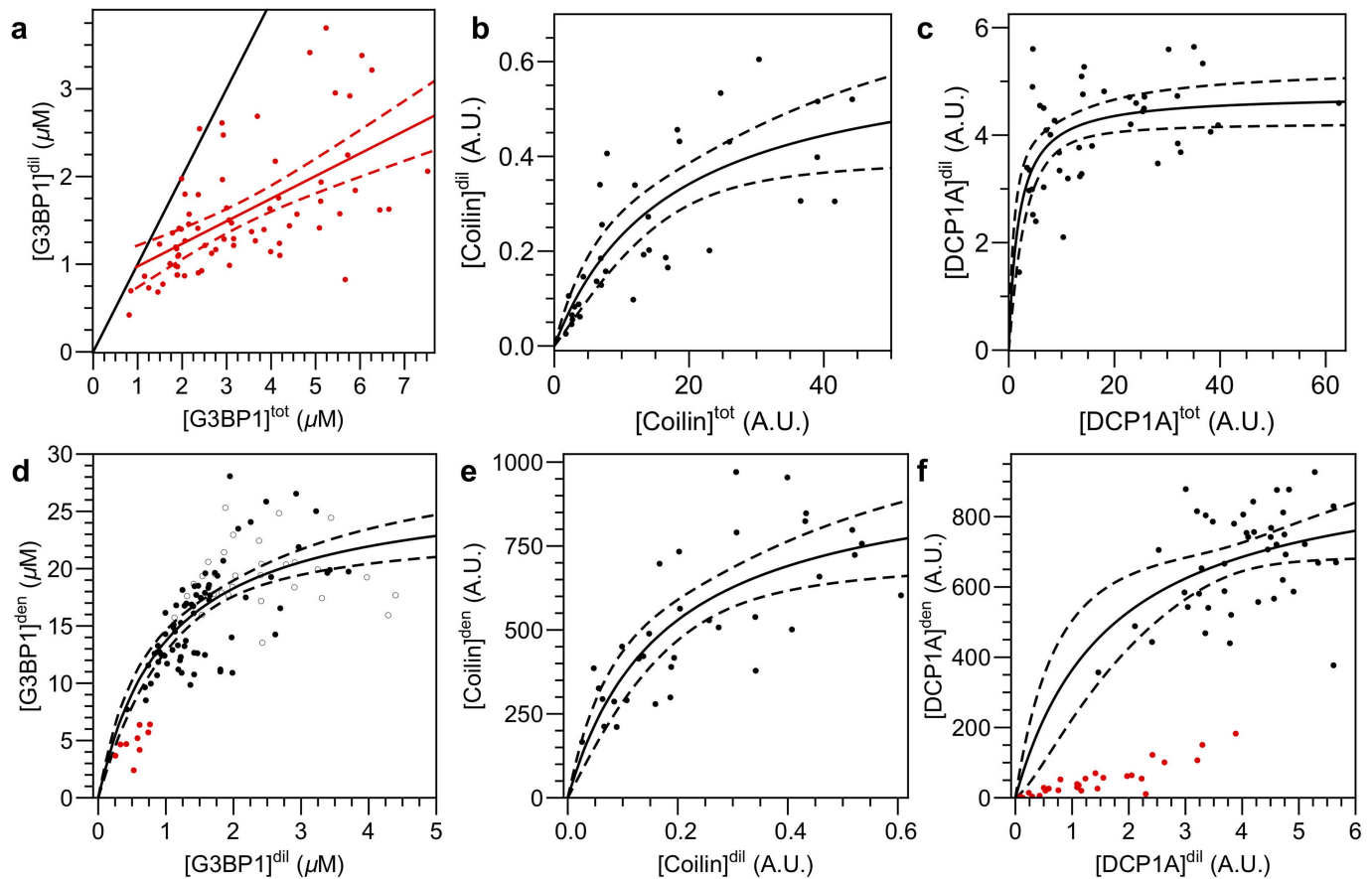
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | NPM1 lacks a fixed C^{dil} and C^{den} , suggesting that nucleoli undergo multicomponent-mediated phase separation.

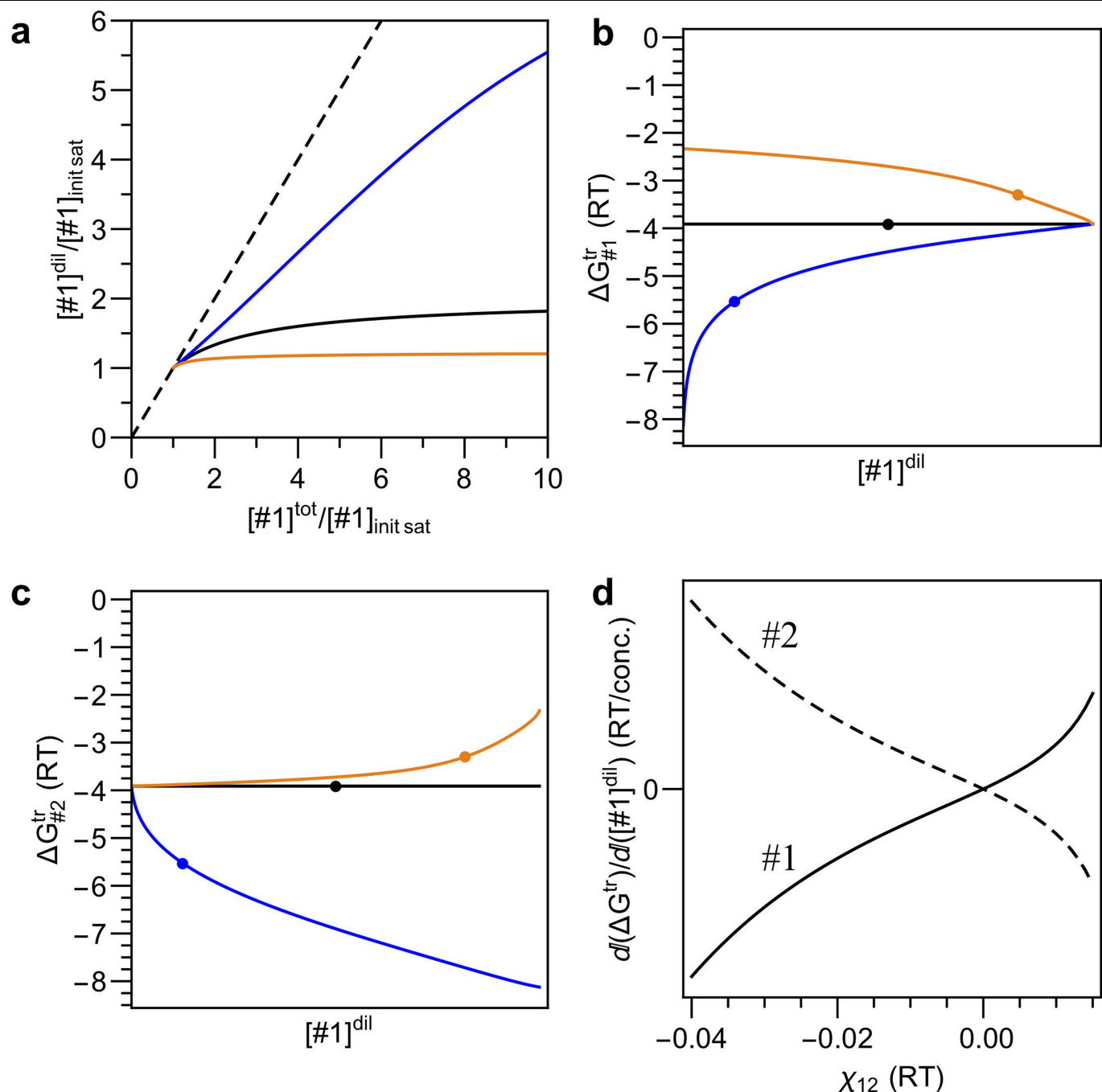
a, b, Dependence of the measured concentration of NPM1 in the relevant dense phase (here 'den' refers to the granular component of nucleoli) (**a**) and the apparent partition coefficient of NPM1 (that is, the ratio of its concentration in the dense and dilute phases) (**b**) on the total concentration of NPM1 in the

nucleus. **c,** Dependence of the transfer free energy on the concentration of NPM1 in the dilute phase, for mCherry-tagged NPM1 (filled circles) and mGFP-tagged (open circles). The trends for each are similar. Dashed lines represent mean confidence intervals to fits described in the Supplementary Methods; the red lines in **a, b** represent expected trends for single-biomolecule phase separation.



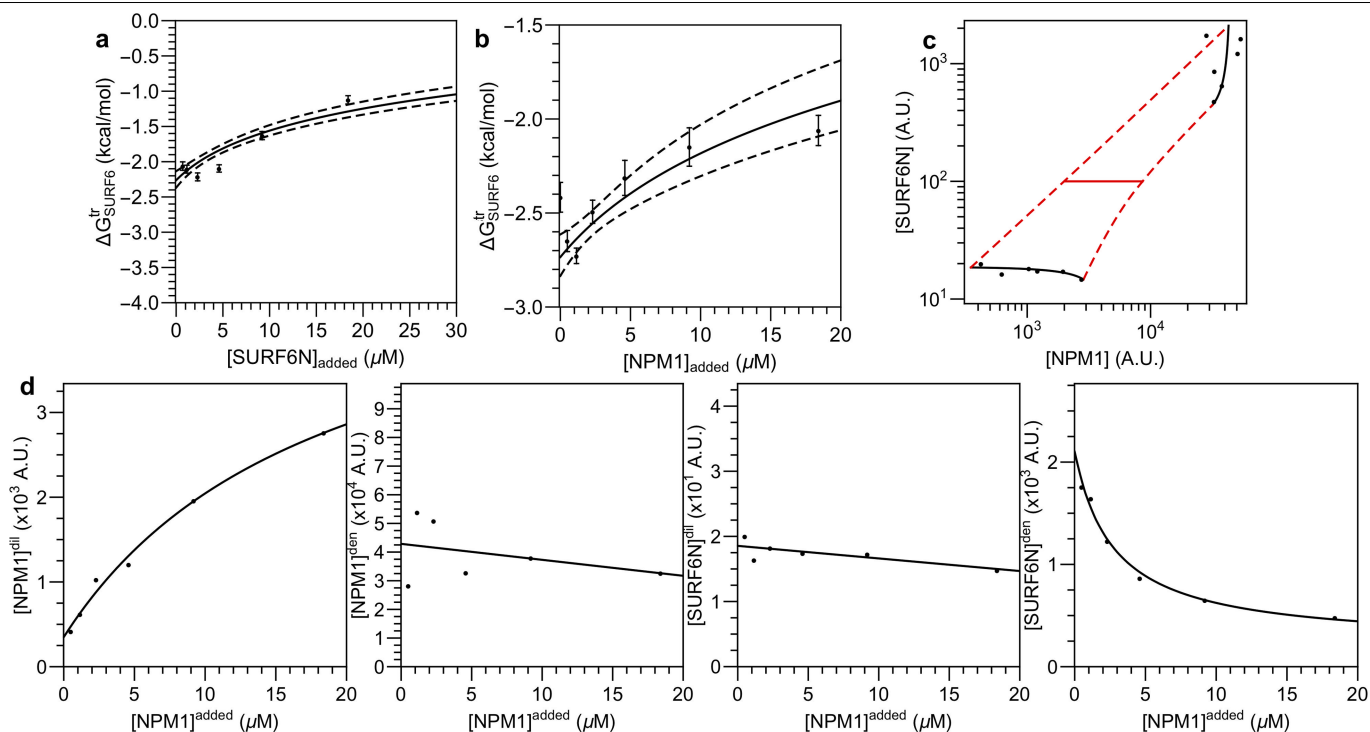
Extended Data Fig. 2 | G3BP1, coilin and DCP1A lack fixed C^{dil} and C^{den} in cells. **a–c**, Relationship between the approximated total concentration and the dilute concentration in cells expressing variable amounts of fluorescently tagged G3BP1 (**a**), coilin (**b**) and DCP1A (**c**). Points are red in **a** to indicate that only cells with phase separation in the G3BP1 double knockout line after stress are included. **d–f**, Relationship between the dilute and dense concentrations

for cells expressing variable amounts of fluorescently tagged G3BP1 (**d**), coilin (**e**) and DCP1A (**f**). Dashed lines represent mean confidence intervals to fits described in the Supplementary Methods. Statistical significance ($P < 0.01$) for these increasing monotonic relationships between the axes are reported in Supplementary Methods. Red points in **d** and **f** represent diffraction-limited foci.



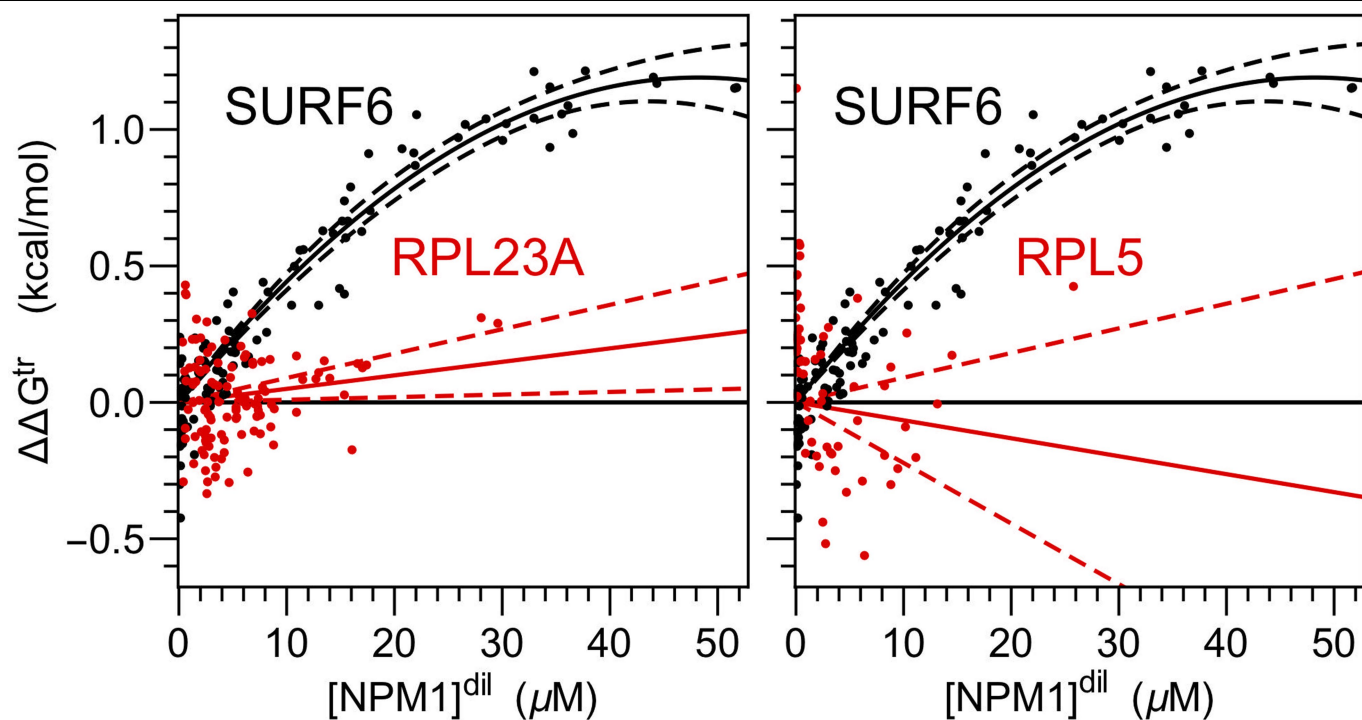
Extended Data Fig. 3 | In silico validation of the composition dependence of phase separation using Flory-Huggins theory. Phase separation of two (non-solvent) components, denoted #1 and #2, with their heterotypic interactions being equal, stronger and weaker, than their homotypic interactions shown as black, blue and orange, respectively, for **a–c**. **a**, The initial dependence of $[\#1]^{dil}$ on $[\#1]^{tot}$ at fixed $[\#2]^{tot}$, such that phase separation will occur at the ‘goldilocks point’—when $[\#1]^{tot} = [\#2]^{tot}$. The axes are normalized by the initial saturation (init sat) concentration—that is, the lowest

$[\#1]^{tot}$ at which phase separation emerges. The dashed line is the 1:1 line that would be expected without phase separation. **b, c**, $\Delta G_{\#1}^{tr}$ (**b**) and $\Delta G_{\#2}^{tr}$ (**c**) as a function of $[\#1]^{dil}$. Circles indicate the location of the goldilocks point under each condition. **d**, The change in ΔG^{tr} with respect to $[\#1]^{dil}$ as a function of the heterotypic interaction strength χ_{12} (in which more negative implies stronger heterotypic interactions) at the goldilocks point for the transfer free energy of #1 and #2, as indicated.

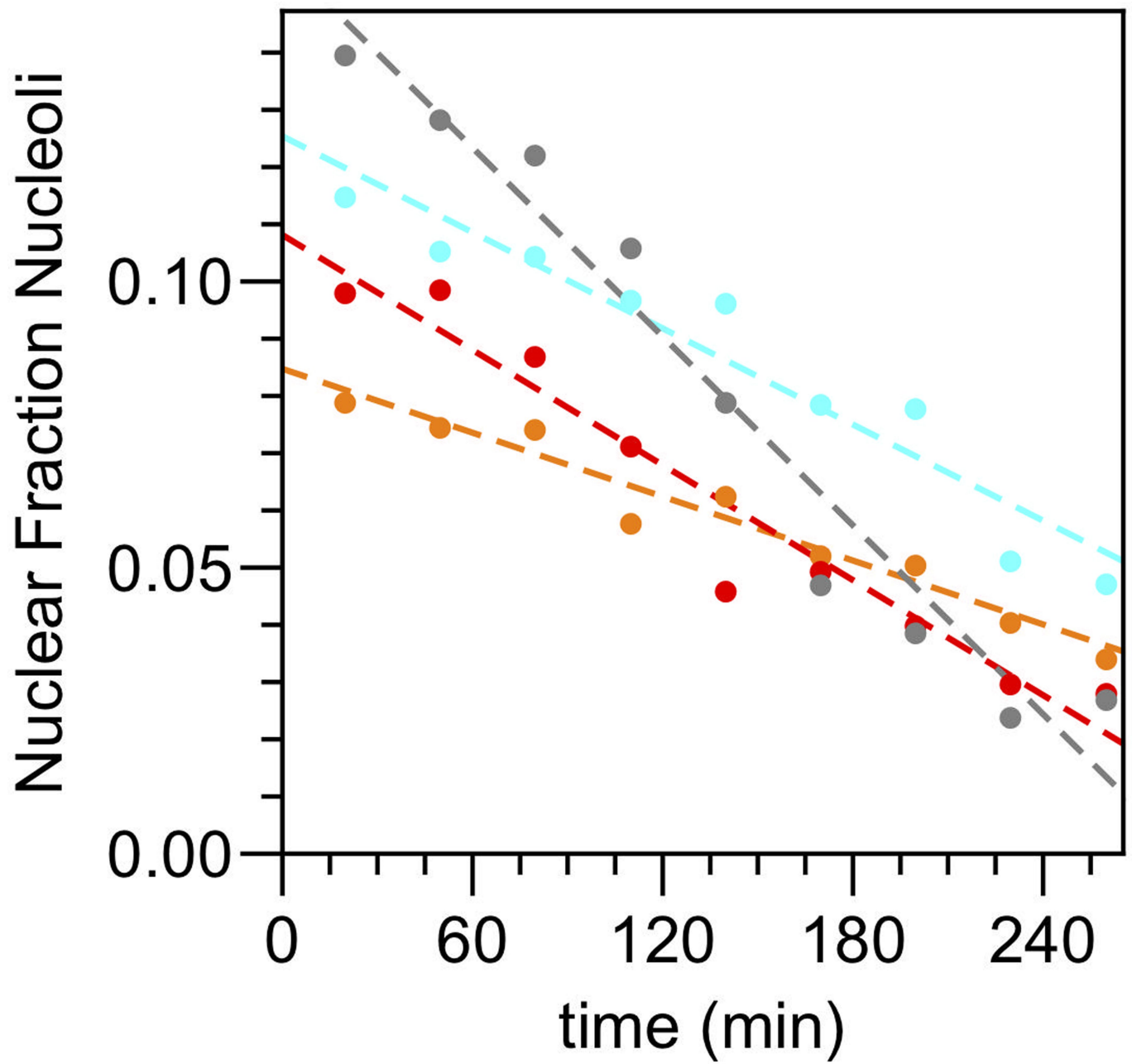


Extended Data Fig. 4 | In vitro destabilization of SURF6N partitioning by increasing the concentration of SURF6N itself or NPM1. **a, b.** Changes in the transfer free energy of SURF6N forming into multicomponent droplets as additional SURF6N (**a**) or additional NPM1 (**b**) is added on top of NPM1:SURF6N:rRNA ternary droplets as described in the Supplementary Methods. The number of droplets (in order of increasing concentration) are $n = 122, 115, 105, 98, 91, 74$ and 99 . Data are mean \pm s.d. **c.** Phase diagram in vitro in the presence of $25 \text{ ng } \mu\text{L}^{-1}$ wheatgerm rRNA, $5 \mu\text{M}$ SURF6N, and various concentrations of NPM1. Units shown are absorbance units corrected for

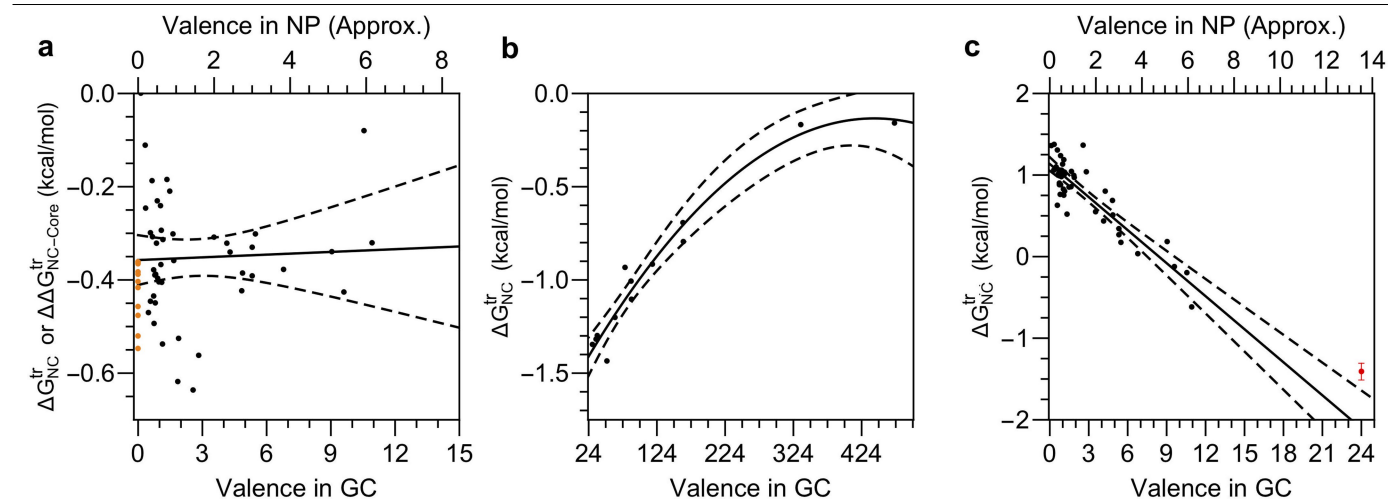
background, quantum yield differences between the two phases, and the (nonlinear) fraction labelled of NPM1. **d.** Changes in the phase diagram as additional NPM1 is added. As in **c**, NPM1 concentrations in the dense or dilute phases are indicative of total NPM1. Hyperbolic fits shown highlight that the largest changes upon NPM1 addition are from an increase in the dilute phase concentration of NPM1 and a decrease in the dense phase concentration of SURF6N. To assess significance, the y axes in **d** are shown from zero arbitrary units (AU) to 2.5 times the mean of all points shown.



Extended Data Fig. 5 | The change in the transfer free energy for R-proteins and NPM1. $\Delta\Delta G^{\text{tr}}$ of r-proteins RPL23A (left) and RPL5 (right) compared with that for SURF6 as the concentration of NPM1 is increased.

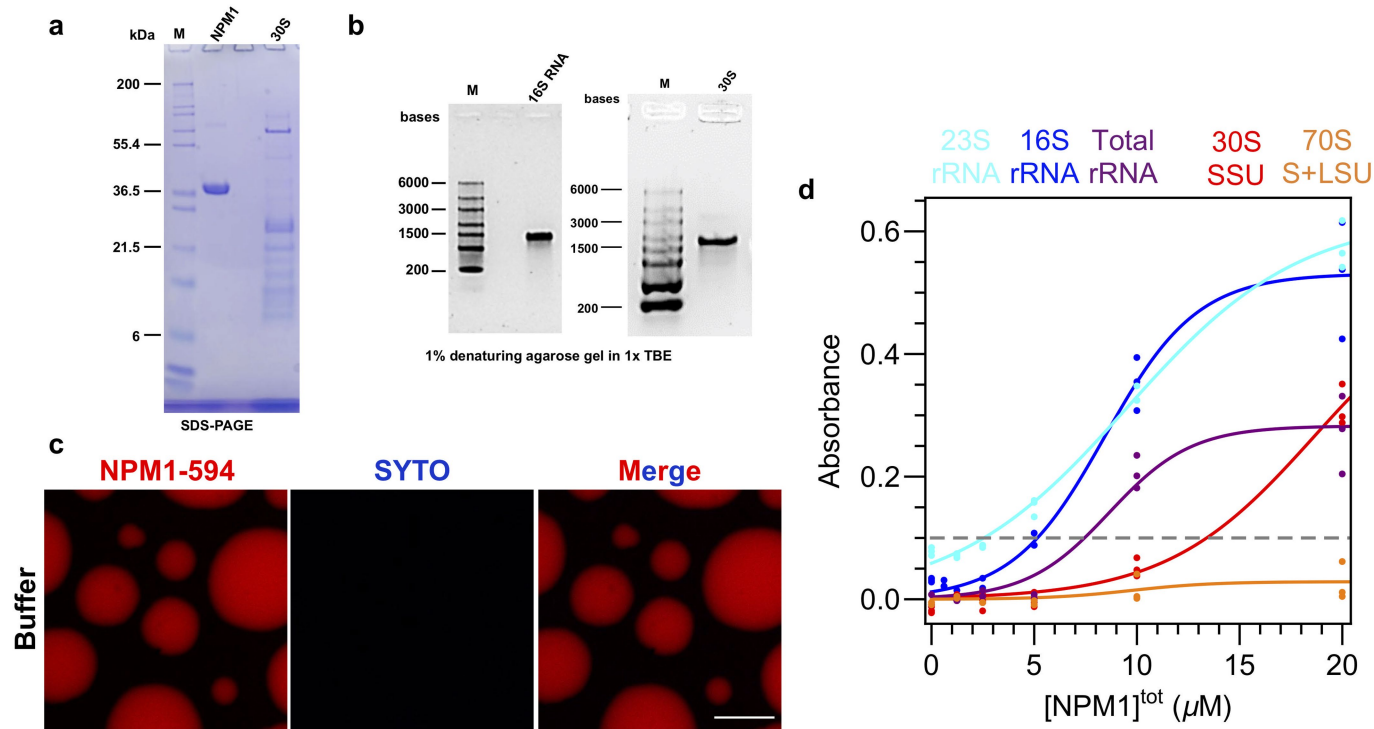


Extended Data Fig. 6 | ActD treatment decreases the size of the nucleolus. The fraction of the image area corresponding to the nucleolus as a function of time after the addition of ActD in individual cells expressing NPM1-mCherry. The colours represent the same cells as in Fig. 3g.



Extended Data Fig. 7 | Characterization of Corelet non-ideality and extrapolation from high valence. **a**, The transfer free energy for the N-terminal half of NPM1 (NC)-sspB in cells without the core expressed (orange representing ΔG^{tr}) or with the indicated valences after core activation (black indicating $\Delta\Delta G_{NC-Core}^{tr}$). The $\Delta\Delta G_{NC-Core}^{tr}$ in this case is the energetic difference between the NC and core channels, which is approximately the energetic difference for transferring an additional NC to the core at that valence. **b**, At valences higher than 24, the transfer free energy is approximated as quadratic

and extrapolated back to a valence of 24 to obtain the transfer free energy at this valence. **c**, Transfer free energy reported from the sspB channel as a function of valence, which is weighted by the number of sspB molecules (owing to the number of mCherry molecules observed being proportional to the valence of each molecule, as opposed to at the core where it is always constant at 24 GFPs). The red point represents the extrapolated value and mean confidence error as determined in **b**.



Extended Data Fig. 8 | Controls for ribosomal mimics. **a, b**, SDS-PAGE (**a**) and denaturing agarose gel (**b**) detailing the purity of reagents used in the experiments in Fig. 4d, e. **c**, Microscopy image of 10 μ M NPM1-594 droplets formed with 5% PEG without any rRNA. The limited fluorescence indicates that neither NPM1 nor PEG binds SYTO 40 and the droplet environment does not

promote the fluorescence of SYTO 40. **d**, Phase separation assessed by turbidity of the indicated ribosomal substrate (fixed at 50 μ g ml⁻¹) as a function of NPM1 concentration. The dashed grey line indicates where phase separation is typically observed in microscopy measurements.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|---|
| n/a | Confirmed |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Source data for Figures 1-4 and Extended Data Figures 1, 2, 3-8 are available with the paper. All other data are available from the corresponding authors upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	At least two dozen or more data points were collected across the largest experimentally possible dynamic range consistent with ideal practice for biophysical research.
Data exclusions	No data were excluded from the analyses except when condensate was below the diffraction limit. This exclusion was predetermined as concentrations of objects below the diffraction limit are immeasurable in standard confocal microscopy.
Replication	All attempts at replication were successful.
Randomization	No randomization was performed as experiments within the study have been highly reproducible (e.g. day to day, locations used on multi-well plate) and often additional data points are collected as a consequence of the previous data (e.g. cells failed to express enough protein).
Blinding	Investigators were not blinded as analysis is required to assure expression levels span required dynamic range as is typical for biophysical experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	HeLa cells were obtained from ATCC. U2OS G3BP1/2 KO cells were previously described (Kedersha et al., 2016).
Authentication	None of the cell lines used were authenticated.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None.

Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor

<https://doi.org/10.1038/s41586-020-2180-5>

Received: 19 February 2020

Accepted: 19 March 2020

Published online: 30 March 2020

 Check for updates

Jun Lan^{1,4}, Jiwan Ge^{1,4}, Jinfang Yu^{1,4}, Sisi Shan^{2,4}, Huan Zhou³, Shilong Fan¹, Qi Zhang², Xuanling Shi², Qisheng Wang³, Linqi Zhang^{2✉} & Xinqun Wang^{1✉}

A new and highly pathogenic coronavirus (severe acute respiratory syndrome coronavirus-2, SARS-CoV-2) caused an outbreak in Wuhan city, Hubei province, China, starting from December 2019 that quickly spread nationwide and to other countries around the world^{1–3}. Here, to better understand the initial step of infection at an atomic level, we determined the crystal structure of the receptor-binding domain (RBD) of the spike protein of SARS-CoV-2 bound to the cell receptor ACE2. The overall ACE2-binding mode of the SARS-CoV-2 RBD is nearly identical to that of the SARS-CoV RBD, which also uses ACE2 as the cell receptor⁴. Structural analysis identified residues in the SARS-CoV-2 RBD that are essential for ACE2 binding, the majority of which either are highly conserved or share similar side chain properties with those in the SARS-CoV RBD. Such similarity in structure and sequence strongly indicate convergent evolution between the SARS-CoV-2 and SARS-CoV RBDs for improved binding to ACE2, although SARS-CoV-2 does not cluster within SARS and SARS-related coronaviruses^{1–3,5}. The epitopes of two SARS-CoV antibodies that target the RBD are also analysed for binding to the SARS-CoV-2 RBD, providing insights into the future identification of cross-reactive antibodies.

The emergence of the highly pathogenic coronavirus SARS-CoV-2 in Wuhan and its rapid international spread has posed a serious global public-health emergency^{1–3}. Similar to individuals who were infected by pathogenic SARS-CoV in 2003 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012, patients infected by SARS-CoV-2 showed a range of symptoms including dry cough, fever, headache, dyspnoea and pneumonia with an estimated mortality rate ranging from 3 to 5%^{6–8}. Since the initial outbreak in December of 2019, SARS-CoV-2 has spread throughout China and to more than 80 other countries and areas worldwide. As of 5 March 2020, 80,565 cases in China have been confirmed with the infection and 3,015 infected patients have died (<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>). As a result, the epicentre Wuhan and the neighbouring cities have been under lockdown to minimize the continued spread and the WHO (World Health Organization) has announced a Public Health Emergency of International Concern owing to the rapid and global dissemination of SARS-CoV-2.

Phylogenetic analyses of the coronavirus genomes have revealed that SARS-CoV-2 is a member of the *Betacoronavirus* genus, which includes SARS-CoV, MERS-CoV, bat SARS-related coronaviruses (SARSr-CoV), as well as others identified in humans and diverse

animal species^{1–3,5}. Bat coronavirus RaTG13 appears to be the closest relative of the SARS-CoV-2, sharing more than 93.1% sequence identity in the spike (S) gene. SARS-CoV and other SARSr-CoVs, however, are distinct from SARS-CoV-2 and share less than 80% sequence identity¹.

Coronaviruses use the homotrimeric spike glycoprotein (comprising a S1 subunit and S2 subunit in each spike monomer) on the envelope to bind to their cellular receptors. Such binding triggers a cascade of events that leads to the fusion between cell and viral membranes for cell entry. Previous cryo-electron microscopy studies of the SARS-CoV spike protein and its interaction with the cell receptor ACE2 have shown that receptor binding induces the dissociation of the S1 with ACE2, prompting the S2 to transit from a metastable pre-fusion to a more-stable post-fusion state that is essential for membrane fusion^{9–12}. Therefore, binding to the ACE2 receptor is a critical initial step for SARS-CoV to enter into target cells. Recent studies also highlighted the important role of ACE2 in mediating entry of SARS-CoV-2^{13–15}. HeLa cells expressing ACE2 are susceptible to SARS-CoV-2 infection whereas those without ACE2 are not¹. In vitro binding measurements also showed that the SARS-CoV-2 RBD binds to ACE2 with an affinity in the low nanomolar range, indicating that the RBD is a key functional component within the S1 subunit that is responsible for binding of SARS-CoV-2 by ACE2^{13,16}.

¹The Ministry of Education Key Laboratory of Protein Science, Beijing Advanced Innovation Center for Structural Biology, Beijing Frontier Research Center for Biological Structure, Collaborative Innovation Center for Biotherapy, School of Life Sciences, Tsinghua University, Beijing, China. ²Center for Global Health and Infectious Diseases, Comprehensive AIDS Research Center, Beijing Advanced Innovation Center for Structural Biology, School of Medicine, Tsinghua University, Beijing, China. ³Shanghai Synchrotron Radiation Facility, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China. ⁴These authors contributed equally: Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan. ✉e-mail: zhanglinqi@mail.tsinghua.edu.cn; xinqunwang@mail.tsinghua.edu.cn

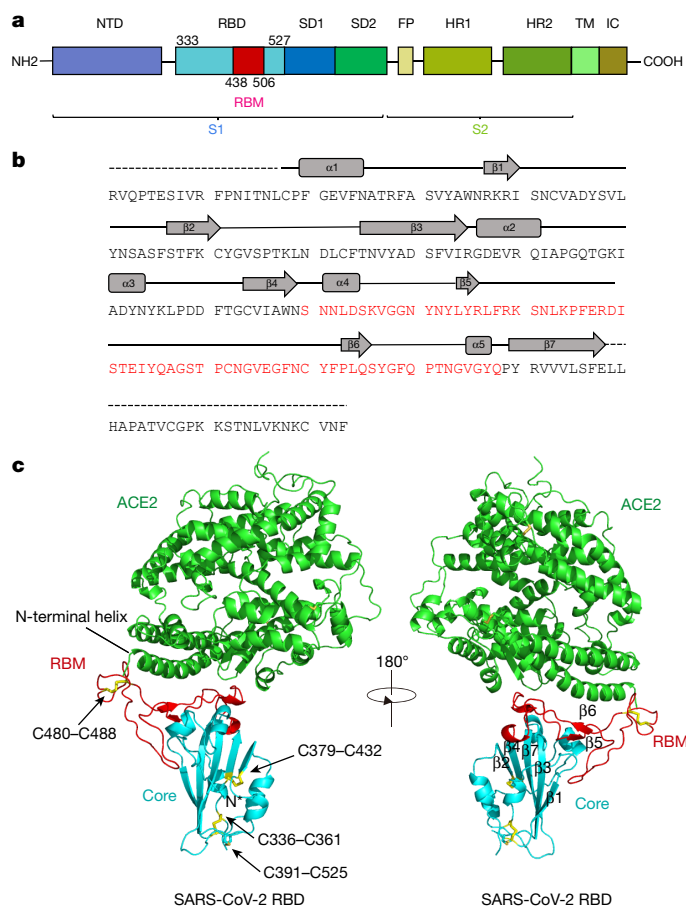


Fig. 1 | Overall structure of SARS-CoV-2 RBD bound to ACE2. **a**, Overall topology of the SARS-CoV-2 spike monomer. FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; IC, intracellular domain; NTD, N-terminal domain; SD1, subdomain 1; SD2, subdomain 2; TM, transmembrane region. **b**, Sequence and secondary structures of SARS-CoV-2 RBD. The RBM sequence is shown in red. **c**, Overall structure of the SARS-CoV-2 RBD bound to ACE2. ACE2 is shown in green. The SARS-CoV-2 RBD core is shown in cyan and RBM in red. Disulfide bonds in the SARS-CoV-2 RBD are shown as sticks and indicated by arrows. The N-terminal helix of ACE2 responsible for binding is labelled.

The cryo-electron microscopy structure of the SARS-CoV-2 spike trimer has recently been reported in two independent studies^{13,17}. However, inspection of one available spike structure revealed the incomplete modelling of the RBD, particularly for the receptor-binding motif (RBM) that interacts directly with ACE2¹⁷. Computer modelling of the interaction between the SARS-CoV-2 RBD and ACE2 has identified some residues that are potentially involved in the interaction; however, the actual residues that mediate the interaction remained unclear¹⁸. Furthermore, despite detectable cross-reactive SARS-CoV-2-neutralizing activity of serum or plasma from patients who recovered from SARS-CoV infections¹⁵, no isolated SARS-CoV monoclonal antibodies are able to neutralize SARS-CoV-2^{16,17}. These findings highlight some of the intrinsic sequence and structure differences between the SARS-CoV and SARS-CoV-2 RBDs.

To elucidate the interaction between the SARS-CoV-2 RBD and ACE2 at a higher resolution, we determined the structure of the SARS-CoV-2 RBD–ACE2 complex using X-ray crystallography. This atomic-level structural information greatly improves our understanding of the interaction between SARS-CoV-2 and susceptible cells, provides a precise target for neutralizing antibodies, and assists the structure-based vaccine design that is urgently needed in the ongoing fight against SARS-CoV-2. Specifically, we expressed

the SARS-CoV-2 RBD (residues Arg319–Phe541) (Fig. 1a, b) and the N-terminal peptidase domain of ACE2 (residues Ser19–Asp615) in Hi5 insect cells and purified them by Ni-NTA affinity purification and gel filtration (Extended Data Fig. 1). The structure of the complex was determined by molecular replacement using the SARS-CoV RBD and ACE2 structures as search models⁴, and refined to a resolution of 2.45 Å with final R_{work} and R_{free} factors of 19.6% and 23.7%, respectively (Extended Data Fig. 2 and Extended Data Table 1). The final model contains residues Thr333–Gly526 of the SARS-CoV-2 RBD, residues Ser19–Asp615 of the ACE2 N-terminal peptidase domain, one zinc ion, four *N*-acetyl- β -glucosaminide (NAG) glycans linked to ACE2 Asn90, Asn322 and Asn546 and to RBD Asn343, as well as 80 water molecules.

The SARS-CoV-2 RBD has a twisted five-stranded antiparallel β sheet (β 1, β 2, β 3, β 4 and β 7) with short connecting helices and loops that form the core (Fig. 1b, c). Between the β 4 and β 7 strands in the core, there is an extended insertion containing the short β 5 and β 6 strands, α 4 and α 5 helices and loops (Fig. 1b, c). This extended insertion is the RBM, which contains most of the contacting residues of SARS-CoV-2 that bind to ACE2. A total of nine cysteine residues are found in the RBD, eight of which form four pairs of disulfide bonds that are resolved in the final model. Among these four pairs, three are in the core (Cys336–Cys361, Cys379–Cys432 and Cys391–Cys525), which help to stabilize the β sheet structure (Fig. 1c); the remaining pair (Cys480–Cys488) connects the loops in the distal end of the RBM (Fig. 1c). The N-terminal peptidase domain of ACE2 has two lobes, forming the peptide substrate binding site between them. The extended RBM in the SARS-CoV-2 RBD contacts the bottom side of the small lobe of ACE2, with a concave outer surface in the RBM that accommodates the N-terminal helix of the ACE2 (Fig. 1c). The overall structure of the SARS-CoV-2 RBD is similar to that of the SARS-CoV RBD (Extended Data Fig. 3a), with a root mean square deviation (r.m.s.d.) of 1.2 Å for 174 aligned C_{α} atoms. Even in the RBM, which has more sequence variation, the overall structure is also highly similar (r.m.s.d. of 1.3 Å) to the SARS-CoV RBD, with only one obvious conformational change in the distal end (Extended Data Fig. 3a). The overall binding mode of the SARS-CoV-2 RBD to ACE2 is also nearly identical to that observed in the previously determined structure of the SARS-CoV RBD–ACE2 complex⁴ (Extended Data Fig. 3b).

The cradling of the N-terminal helix of ACE2 by the outer surface of the RBM results in a large buried surface of 1,687 Å² (864 Å² on the RBD and 823 Å² on the ACE2) at the SARS-CoV-2 RBD–ACE2 interface. A highly similar buried surface of 1,699 Å² contributed by SARS-CoV RBD (869 Å²) and ACE2 (830 Å²) is also observed at the SARS-CoV RBD–ACE2 interface. With a distance cut-off of 4 Å, a total of 17 residues of the RBD are in contact with 20 residues of ACE2 (Fig. 2a and Extended Data Table 2). Analysis of the interface between the SARS-CoV RBD and ACE2 revealed a total of 16 residues of the SARS-CoV RBD in contact with 20 residues of ACE2 (Fig. 2a and Extended Data Table 2). Among the 20 ACE2 residues that interact with the two different RBDs, 17 residues are shared between both interactions and most of the contacting residues are located at the N-terminal helix (Fig. 2a and Extended Data Table 2).

To compare the ACE2-interacting residues on the SARS-CoV-2 and SARS-CoV RBDs, we used structure-guided sequence alignment and mapped them to their respective sequences (Fig. 2b). Among 14 shared amino acid positions used by both RBMs for the interaction with ACE2, 8 have the identical residues between the two RBDs, including Tyr449/Tyr436, Tyr453/Tyr440, Asn487/Asn473, Tyr489/Tyr475, Gly496/Gly482, Thr500/Thr486, Gly502/Gly488 and Tyr505/Tyr491 of SARS-CoV-2/SARS-CoV, respectively (Fig. 2b). Five positions have residues that have similar biochemical properties despite of having different side chains, including Leu455/Tyr442, Phe456/Leu443, Phe486/Leu472, Gln493/Asn479 and Asn501/Thr487 of SARS-CoV-2/SARS-CoV, respectively (Fig. 2b).

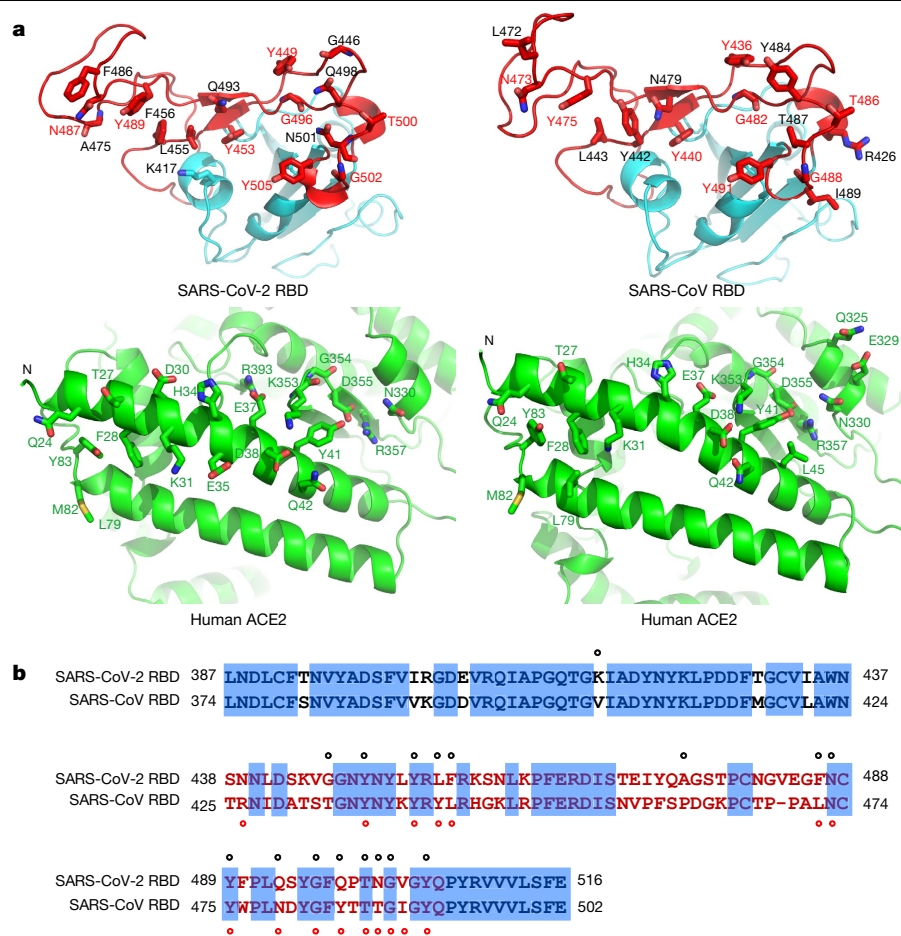


Fig. 2 | The SARS-CoV-2 RBD-ACE2 and SARS-CoV RBD-ACE2 interfaces. **a**, Contacting residues are shown as sticks at the SARS-CoV-2 RBD-ACE2 and SARS-CoV RBD-ACE2 interfaces. Positions in both RBDs that are involved in ACE2 binding are indicated by red labels. **b**, Sequence alignment of the

SARS-CoV-2 and SARS-CoV RBDs. Contacting residues in the SARS-CoV-2 RBD are indicated by black dots; contacting residues in the SARS-CoV RBD are indicated by red dots.

The remaining position is at the Gln498/Tyr484 location (Fig. 2b), at which Gln498 of SARS-CoV-2 and Tyr484 of SARS-CoV both interact with Asp38, Tyr41, Gln42, Leu45 and Lys353 of ACE2. Among the six RBD positions with changed residues, SARS-CoV residues Tyr442, Leu472, Asn479 and Thr487 have previously been shown to be essential for binding ACE2¹⁸. At the Leu455/Tyr442 position, Leu455 of SARS-CoV-2 and Tyr442 of SARS-CoV have similar interactions with Asp30, Lys31 and His34 of ACE2 (Fig. 3a). At the Phe486/Leu472 position, Phe486 of SARS-CoV-2 interacts with Gln24, Leu79, Met82 and Tyr83 of ACE2, whereas Leu472 of SARS-CoV has less interactions with Leu79 and Met82 of ACE2 (Fig. 3a). At the Gln493/Asn479 position, Gln493 of SARS-CoV-2 interacts with Lys31, His34 and Glu35 of ACE2 and forms a hydrogen bond with Glu35; Asn479 of SARS-CoV interacts with only His34 of ACE2 (Fig. 3a). At the Asn501/Thr487 position, both residues have similar interactions with Tyr41, Lys353, Gly354 and Asp355 of ACE2 (Fig. 3a). Asn501 of SARS-CoV-2 and Thr487 of SARS-CoV both form a hydrogen bond with Tyr41 of ACE2 (Fig. 3a). Outside the RBM, there is a unique ACE2-interacting residue (Lys417) in SARS-CoV-2, which forms salt-bridge interactions with Asp30 of ACE2 (Fig. 3b). This position is replaced by a valine in the SARS-CoV RBD that fails to participate in ACE2 binding (Figs. 2b, 3b). Furthermore, a comparison of the surface electrostatic potential also identified a positive charged patch on the SARS-CoV-2 RBD contributed by Lys417 that is absent on the SARS-CoV RBD (Fig. 3b). These subtly different ACE2 interactions may contribute to the difference in binding affinity of the SARS-CoV-2 and SARS-CoV

to the ACE2 receptor (4.7 nM compared with 31 nM, respectively) (Extended Data Fig. 4).

One notable and common feature that was found for both RBD-ACE2 interfaces is the networks of hydrophilic interactions. There are 13 hydrogen bonds and 2 salt bridges at the SARS-CoV-2 RBD-ACE2 interface, and 13 hydrogen bonds and 3 salt bridges at the SARS-CoV RBD-ACE2 interface (Table 1). The second shared feature is the involvement of multiple tyrosine residues that form hydrogen-bonding interactions with the polar hydroxyl group. These include Tyr449, Tyr489 and Tyr505 from the SARS-CoV-2 RBD and Tyr436, Tyr475 and Tyr491 from the SARS-CoV RBD (Table 1). The third shared feature may reside in the Asn90-linked glycans of the ACE2 that bind to different RBDs. In the structure of the SARS-CoV RBD-ACE2 complex, a chain of Asn90-linked NAG-NAG- β -D-mannose is in contact with Thr402 of the SARS-CoV RBD (Extended Data Fig. 5a), and this glycan-RBD interaction has been proposed to have important roles in the binding of SARS-CoV RBD by ACE2^{4,19}. In the SARS-CoV-2 RBD-ACE2 structure, the density enabled only the modelling of the first NAG linked to ACE2 Asn90, and no interactions between this NAG and the SARS-CoV-2 RBD were observed (Extended Data Fig. 5b). However, this does not exclude that glycans after the first NAG may interact with the SARS-CoV-2 RBD and may have important roles in the binding of SARS-CoV-2 RBD by ACE2. Taken together, our results show that the SARS-CoV-2 RBD-ACE2 and SARS-CoV RBD-ACE2 interfaces share substantial similarity in the buried surface area, the number of interacting residues and hydrophilic interaction networks, although

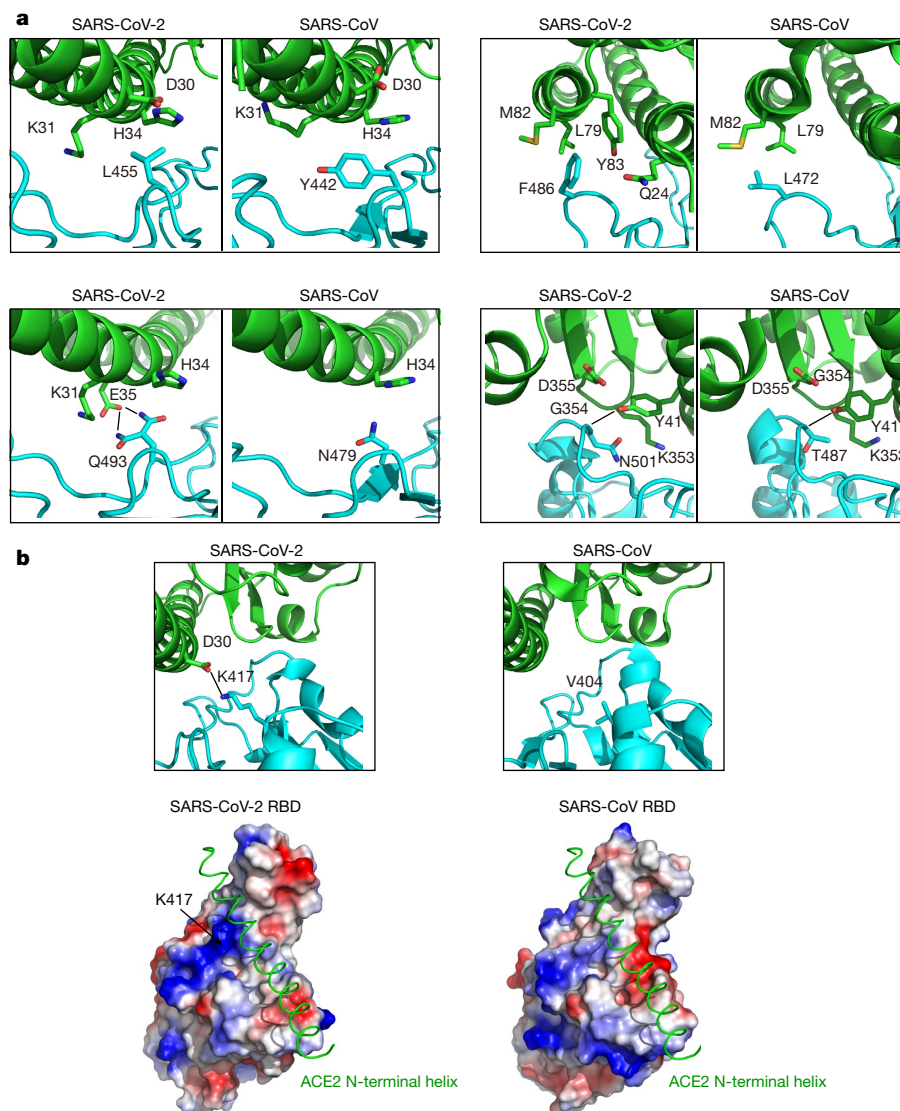


Fig. 3 | Comparisons of interactions at the SARS-CoV-2 RBD-ACE2 and SARS-CoV RBD-ACE2 interfaces. a, Interactions around the SARS-CoV-2 and SARS-CoV positions in the RBM with changed residues. SARS-CoV-2 and SARS-CoV RBDs are shown in cyan. ACE2 is shown in green. **b,** Interactions around the K417 and V404 positions of SARS-CoV-2 and SARS-CoV RBDs,

respectively, that are outside the RBM and electrostatic potential maps of the SARS-CoV-2 and SARS-CoV RBDs. The position of K417 in the SARS-CoV-2 RBD is indicated by a black arrow. The N-terminal helix of ACE2 is shown as a green ribbon. The Protein Data Bank (PDB) code for the SARS-CoV RBD-ACE2 complex is 2AJF.

some of the ACE2 interactions observed both inside and outside the RBM were different (Fig. 3a, b). Such similarities argue strongly for the convergent evolution of the SARS-CoV-2 and SARS-CoV RBD structures to improve binding affinity to the same ACE2 receptor, although SARS-CoV-2 does not cluster within SARS-CoV and SARSr-CoV in the *Betacoronavirus* genus.

Consistent with the high structural similarity, we found that the binding affinities between ACE2 and SARS-CoV-2 and SARS-CoV RBDs also fall into a similar range. Specifically, the equilibrium dissociation constant (K_D) of ACE2 and SARS-CoV-2 RBD is 4.7 nM, and of ACE2 and SARS-CoV RBD is 31 nM (Extended Data Fig. 4). Similar results have also been reported by other groups^{13,16}. However, this is slightly different from a recent report in which an approximately 20-fold increased binding between ACE2 and the SARS-CoV-2 spike trimer was found (K_D of 14.7 nM) compared with that between ACE2 and SARS-CoV RBD-SD1 (K_D of 325 nM)¹⁷. This is perhaps due to the different proteins used in the assay or because of other unknown reasons. Nevertheless, the binding affinity alone is unlikely to explain the unusual transmissibility of SARS-CoV-2. Other factors such as the unique 'RRAR' furin cleavage

site at the S1-S2 boundary of the SARS-CoV-2 spike protein may have more-important roles in facilitating the rapid human-to-human transmission of SARS-CoV-2.

Neutralizing antibodies represent an important component of the immune system in the fight against viral infections. It has been reported that SARS-CoV-2 could be cross-neutralized by horse anti-SARS-CoV serum and convalescent serum from a patient with a SARS-CoV infection^{1,15}, reinforcing the structural similarity between the RBDs of SARS-CoV-2 and SARS-CoV. Such similarity also increased the hope of the rapid application of previously characterized SARS-CoV monoclonal antibodies in the clinical setting. However, no antibody that targeted SARS-CoV (m396, S230, 80R and CR3014) has so far demonstrated any notable cross-binding and neutralization activity against spike protein or RBD of SARS-CoV-2^{16,17,20-23}. One exception is SARS-CoV antibody CR3022 that binds to the SARS-CoV-2 RBD with a K_D of 6.2 nM, although its neutralizing activity against SARS-CoV-2 has not yet been reported¹⁶. Currently, we are uncertain where exactly the epitope of CR3022 on the RBDs of SARS-CoV or SARS-CoV-2 is located. Among the three antibodies that are incapable

Table 1 | The hydrogen bonds and salt bridges at the SARS-CoV-2 RBD–ACE2 and SARS-CoV RBD–ACE2 interfaces

	SARS-CoV-2 RBD	Length (Å)	ACE2	Length (Å)	SARS-CoV RBD
Hydrogen bonds	N487(ND2)	2.6	Q24(OE1)	2.9	N473(ND2)
	K417(NZ)	3.0	D30(OD2)		
	Q493(NE2)	2.8	E35(OE2)		
			E37(OE1)	3.4	Y491(OH)
	Y505(OH)	3.2	E37(OE2)		
			D38(OD1)	3.0	Y436(OH)
	Y449(OH)	2.7	D38(OD2)	3.0	Y436(OH)
	T500(OG1)	2.6	Y41(OH)	2.8	T486(OG1)
	N501(N)	3.7	Y41(OH)	3.3	T487(N)
	G446(O)	3.3	Q42(NE2)		
	Y449(OH)	3.0	Q42(NE2)		
			Q42(OE1)	2.7	Y436(OH)
	Y489(OH)	3.5	Y83(OH)	3.3	Y475(OH)
	N487(OD1)	2.7	Y83(OH)	2.8	N473(ND2)
			Q325(OE1)	3.8	R426(NH2)
			E329(OE2)	3.0	R426(NH2)
			N330(ND2)	2.8	T486(O)
	G502(N)	2.8	K353(O)	2.6	G488(N)
	Y505(OH)	3.7	R393(NH2)		
Salt bridges	K417(NZ)	3.9	D30(OD1)		
	K417(NZ)	3.0	D30(OD2)		
			E329(OE2)	3.7	R426(NH1)
			E329(OE1)	3.9	R426(NH2)
			E329(OE2)	3.0	R426(NH2)

ND2, nitrogen delta 2; NE2, nitrogen epsilon 2; NZ, nitrogen zeta; N, nitrogen; NH1, nitrogen eta 1; NH2, nitrogen eta 2; OH, oxygen eta; O, oxygen; OD1, oxygen delta 1; OD2, oxygen delta 2; OG1, oxygen gamma 1; OE1, oxygen epsilon 1; OE2, oxygen epsilon 2.

of binding to the SARS-CoV-2 RBD, two (m396 and 80R) have their epitopes resolved by the high-resolution crystal-structure determination of SARS-CoV RBD–Fab complexes^{20,21}. By mapping these epitope residues onto the sequence of SARS-CoV RBD aligned with the sequence of SARS-CoV-2 RBD (Fig. 4), we found that antibody m396 has 7 residue changes in the SARS-CoV-2 RBD among 21 epitope positions (Fig. 4). There are 16 residue changes in the SARS-CoV-2 RBD among 25 epitope positions of antibody 80R (Fig. 4). This may provide a structural basis for the lack of cross-reactivity of m396 and 80R with SARS-CoV-2. The cross-neutralization of SARS-CoV-2 by horse anti-SARS-CoV serum and serum or plasma from patients recovered from SARS-CoV infections reveals a great potential in

identifying antibodies with cross-reactivity between these two coronaviruses^{1,15}. The conserved non-RBD regions in the spike protein, such as the S2 subunit, are the potential targets for cross-reactive antibodies. Although the RBD is less conserved, identical residues between SARS-CoV-2 and SARS-CoV RBD exist, even in the more variable RBM (Fig. 4). Considering that the RBD is the important region for receptor binding, antibodies that target the conserved epitopes in the RBD will also present a great potential for developing highly potent cross-reactive therapeutic agents against diverse coronavirus species, including SARS-CoV-2.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2180-5>.

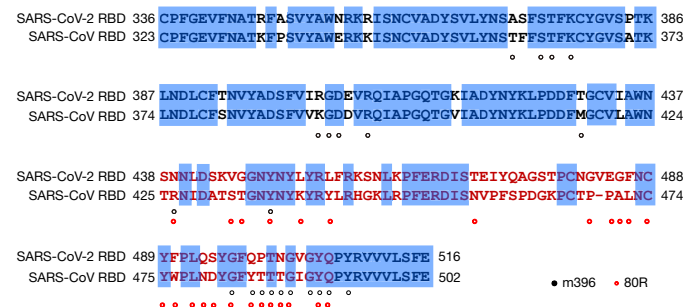


Fig. 4 | Mapping of SARS-CoV neutralizing antibody epitopes. The epitopes of SARS-CoV neutralizing antibodies m396 and 80R, which target the RBD, are labelled in the SARS-CoV sequence aligned with the sequence of SARS-CoV-2 RBD. Epitope residues of m396 are indicated by black dots; epitope residues of 80R are indicated by red dots.

1. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
2. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
3. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
4. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864–1868 (2005).
5. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
6. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
7. Liu, K. et al. Clinical characteristics of novel coronavirus cases in tertiary hospitals in Hubei Province. *Chin. Med. J. (Engl.)* <https://doi.org/10.1097/CM9.0000000000000744> (2020).

8. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *J. Am. Med. Assoc.* **323**, 1061–1069 (2020).
9. Gui, M. et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res.* **27**, 119–129 (2017).
10. Song, W., Gui, M., Wang, X. & Xiang, Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog.* **14**, e1007236 (2018).
11. Kirchdoerfer, R. N. et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci. Rep.* **8**, 15701 (2018).
12. Yuan, Y. et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat. Commun.* **8**, 15092 (2017).
13. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* <https://doi.org/10.1016/j.cell.2020.02.058> (2020).
14. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
15. Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* <https://doi.org/10.1016/j.cell.2020.02.052> (2020).
16. Tian, X. et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg. Microbes Infect.* **9**, 382–385 (2020).
17. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
18. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS. *J. Virol.* **94**, e00127-20 (2020).
19. Li, W. et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.* **24**, 1634–1643 (2005).
20. Prabakaran, P. et al. Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. *J. Biol. Chem.* **281**, 15829–15836 (2006).
21. Hwang, W. C. et al. Structural basis of neutralization by a human anti-severe acute respiratory syndrome spike protein antibody, 80R. *J. Biol. Chem.* **281**, 34610–34616 (2006).
22. Walls, A. C. et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell* **176**, 1026–1039 (2019).
23. van den Brink, E. N. et al. Molecular and biological characterization of human monoclonal antibodies binding to the spike and nucleocapsid proteins of severe acute respiratory syndrome coronavirus. *J. Virol.* **79**, 1635–1644 (2005).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification

The SARS-CoV-2 RBD and the N-terminal peptidase domain of human ACE2 were expressed using the Bac-to-Bac baculovirus system (Invitrogen). The SARS-CoV-2 RBD (residues Arg319–Phe541) with an N-terminal gp67 signal peptide for secretion and a C-terminal 6×His tag for purification was inserted into the pFastBac-Dual vector (Invitrogen). The construct was transformed into bacterial DH10Bac competent cells, and the extracted bacmid was then transfected into Sf9 cells using Cellfectin II Reagent (Invitrogen). The low-titre viruses were collected and then amplified to generate high-titre virus stocks, which were used to infect Hi5 cells at a density of 2×10^6 cells per ml. The supernatant of cell culture containing the secreted SARS-CoV-2 RBD was collected 60 h after infection, concentrated and buffer-exchanged to HBS (10 mM HEPES, pH 7.2, 150 mM NaCl). The SARS-CoV-2 RBD was captured by Ni-NTA resin (GE Healthcare) and eluted with 500 mM imidazole in HBS buffer. The SARS-CoV-2 RBD was then purified by gel filtration chromatography using a Superdex 200 column (GE Healthcare) pre-equilibrated with HBS buffer. Fractions containing the SARS-CoV-2 RBD were collected.

The N-terminal peptidase domain of human ACE2 (residues Ser19–Asp615) was expressed and purified by essentially the same protocol as used for the SARS-CoV-2 RBD. To purify the SARS-CoV-2 RBD–ACE2 complex, ACE2 was incubated with the SARS-CoV-2 RBD for 1 h on ice in HBS buffer, and the mixture was then subjected to gel filtration chromatography. Fractions containing the complex were pooled and concentrated to 13 mg ml⁻¹.

Crystallization and data collection

Crystals were successfully grown at room temperature in sitting drops, over wells containing 100 mM MES, pH 6.5, 10% PEG 5000 MME and 12% 1-propanol. The drops were made by mixing 200 nl of the SARS-CoV-2 RBD–ACE2 complex in 20 mM Tris pH 7.5, 150 mM NaCl with 200 nl well solution. Crystals were collected, soaked briefly in 100 mM MES, pH 6.5, 10% PEG 5000 MME, 12% 1-propanol and 20% glycerol, and were subsequently flash-frozen in liquid nitrogen. Diffraction data were collected at 100 K and at a wavelength of 1.07180 Å on the BL17U1 beam line of the Shanghai Synchrotron Research Facility. Diffraction data were autoprocessed using the aquarium pipeline²⁴ and the data-processing statistics are listed in Extended Data Table 1.

Structure determination and refinement

The structure was determined using the molecular replacement method with PHASER in the CCP4 suite²⁵. The search models used included the ACE2 extracellular domain and SARS-CoV RBD (PDB code 2AJF). Density map improvement by updating and refinement of the atoms was performed with ARP/wARP²⁶. Subsequent model building and refinement were performed using COOT and PHENIX, respectively^{27,28}. Final Ramachandran statistics: 96.44% favoured, 3.56% allowed and 0.00%

outliers for the final structure. The structure refinement statistics are listed in Extended Data Table 1. All structure figures were generated with PyMol²⁹.

Surface plasmon resonance experiments

ACE2 was immobilized on a CM5 sensorchip (GE Healthcare) to a level of around 500 response units using a Biacore T200 (GE Healthcare) and a running buffer composed of 10 mM HEPES pH 7.2, 150 mM NaCl and 0.05% Tween-20. Serial dilutions of the SARS-CoV RBD and SARS-CoV-2 RBD were flowed through with a concentration ranging from 62.5 to 1.9 nM. The resulting data were fit to a 1:1 binding model using Biacore Evaluation Software (GE Healthcare).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The coordinates and structure factor files for the SARS-CoV-2 RBD–ACE2 complex have been deposited in the Protein Data Bank (PDB) under accession number 6M0J.

24. Yu, F. et al. Aquarium: an automatic data-processing and experiment information management system for biological macromolecular crystallography beamlines. *J. Appl. Crystallogr.* **52**, 472–477 (2019).
25. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
26. Cohen, S. X. et al. ARP/wARP and molecular replacement: the next generation. *Acta Crystallogr. D* **64**, 49–60 (2008).
27. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
28. Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
29. Janson, G., Zhang, C., Prado, M. G. & Paiardini, A. PyMod 2.0: improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics* **33**, 444–446 (2017).

Acknowledgements We thank the staff at the BL17U1 beam line of the Shanghai Synchrotron Research Facility for data collection and processing. We thank staff at the X-ray crystallography platform of the Tsinghua University Technology Center for Protein Research for providing facility support. This work was supported by funds from the National Key Plan for Scientific Research and Development of China (grant number 2016YFD0500307). Research is also supported by the Tsinghua University Initiative Scientific Research Program (20201080053), National Natural Science Foundation Award (81530065), Beijing Municipal Science and Technology Commission (171100000517-001 and 171100000517-003), Tencent Foundation, Shuidi Foundation and TH Capital.

Author contributions J.L. and J.G. carried out protein expression, purification, crystallization, diffraction data collection and structure determination with the help of J.Y. and S.S. Q.Z. and X.S. helped with protein expression and purification. S.F., H.Z. and Q.W. helped with the collection of crystallization and diffraction data. X.W. and L.Z. conceived, designed and directed the study. J.L., J.G., J.Y., S.S., L.Z. and X.W. analysed the data, generated the figures and wrote the manuscript.

Competing interests The authors declare no competing interests.

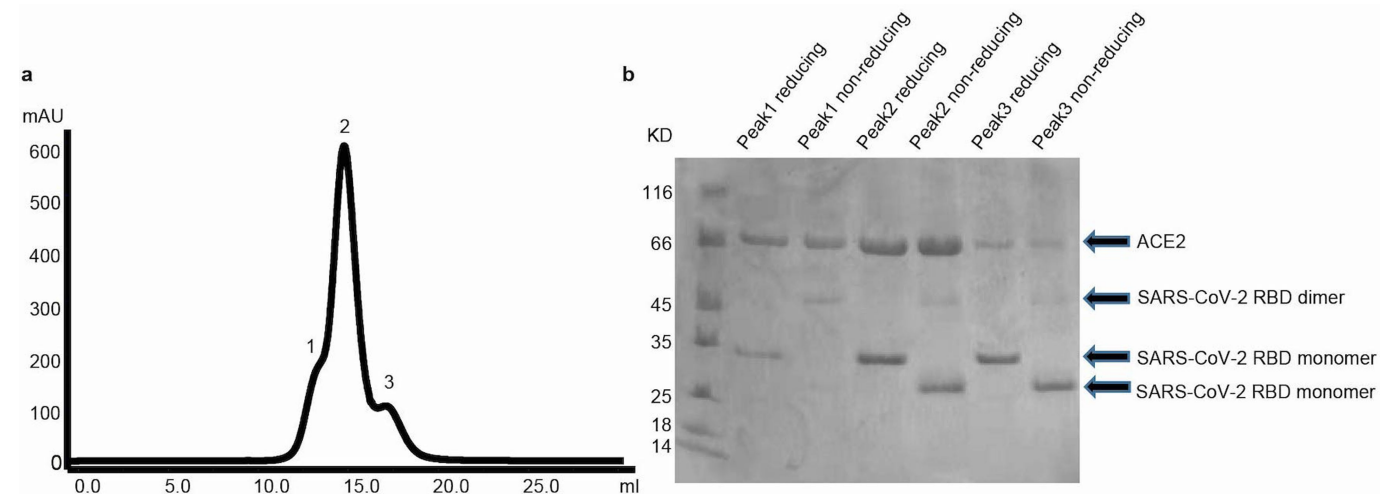
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2180-5>.

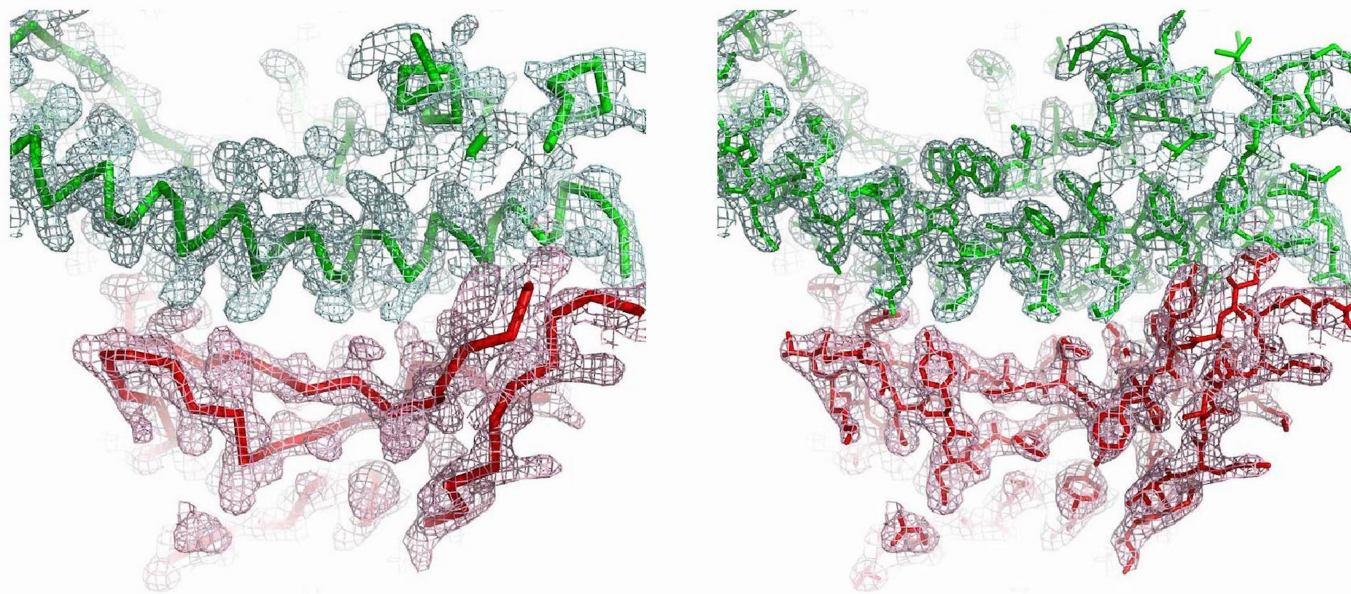
Correspondence and requests for materials should be addressed to L.Z. or X.W.

Peer review information Nature thanks Lijun Rong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

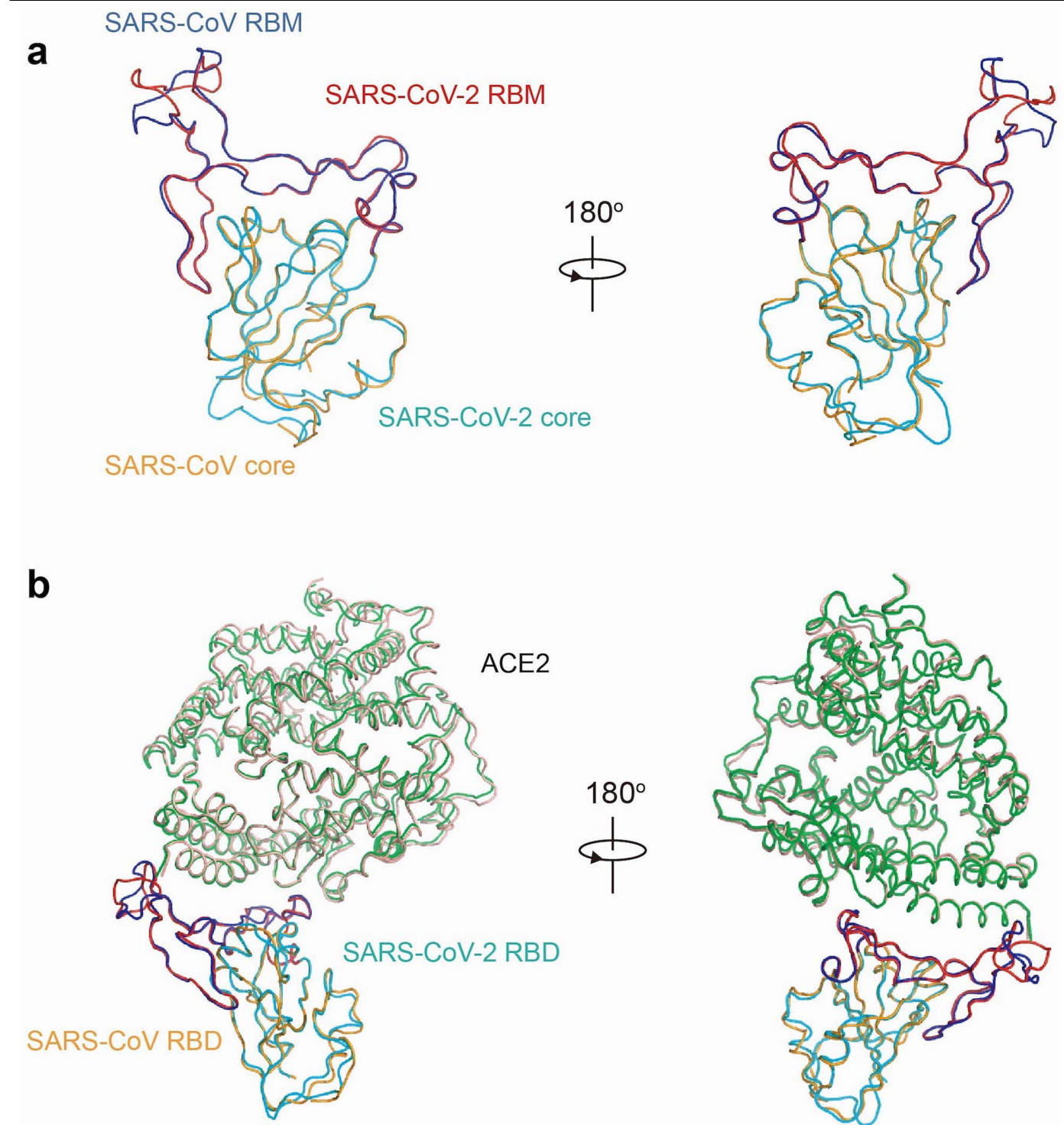
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Purification of the SARS-CoV-2 RBD-ACE2 complex. **a**, Gel filtration chromatography of the complex. 1, SARS-CoV-2 RBD dimer-ACE2; 2, SARS-CoV-2 RBD monomer-ACE2; 3, SARS-CoV-2 RBD monomer. **b**, SDS-PAGE gel of the complex.

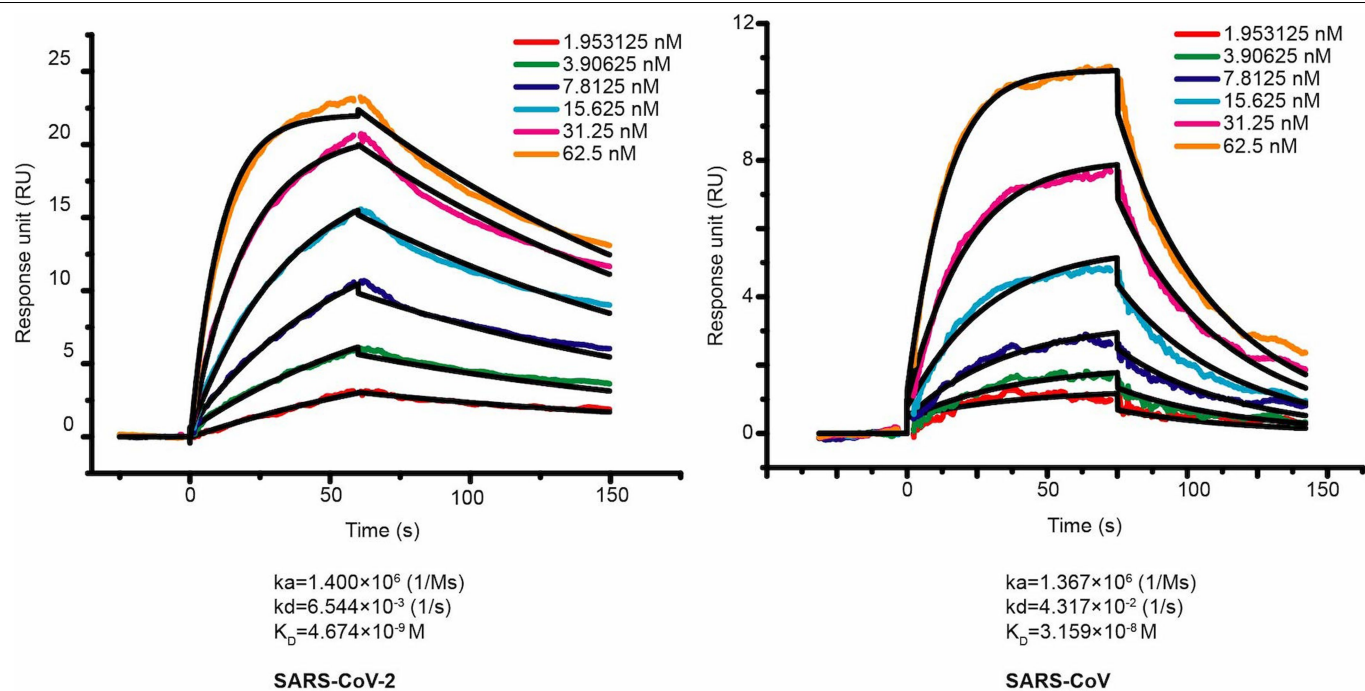


Extended Data Fig. 2 | Electron density map. $2F_o - F_c$ electron density maps contoured at 1.5σ at the binding interface between the SARS-CoV-2 RBD (red) and ACE2 (green) are shown.

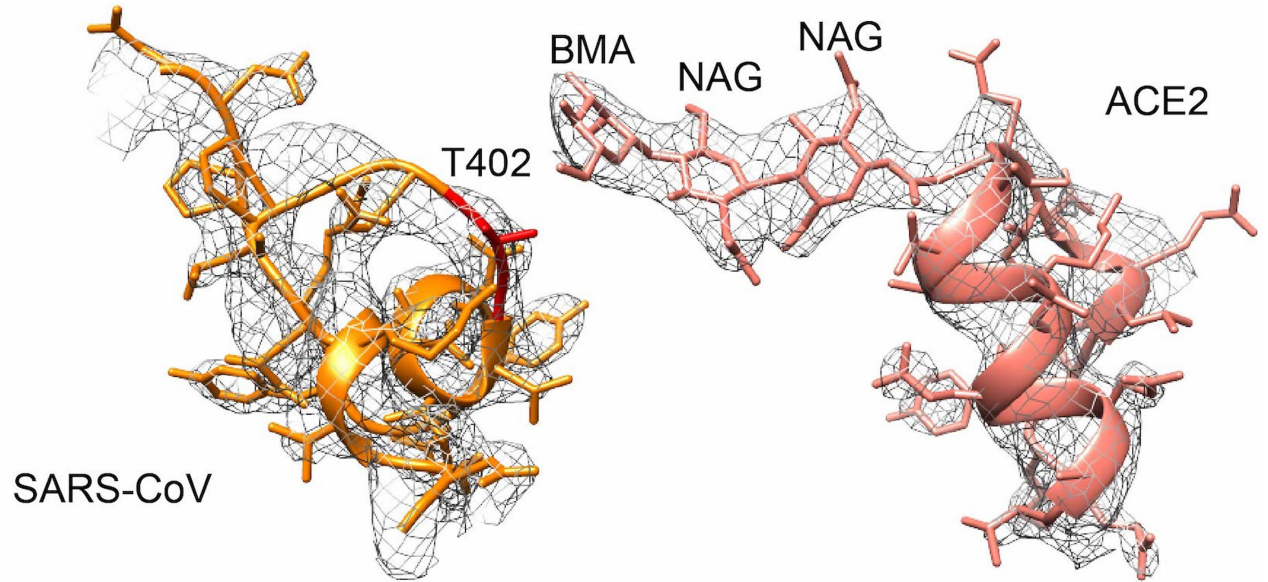
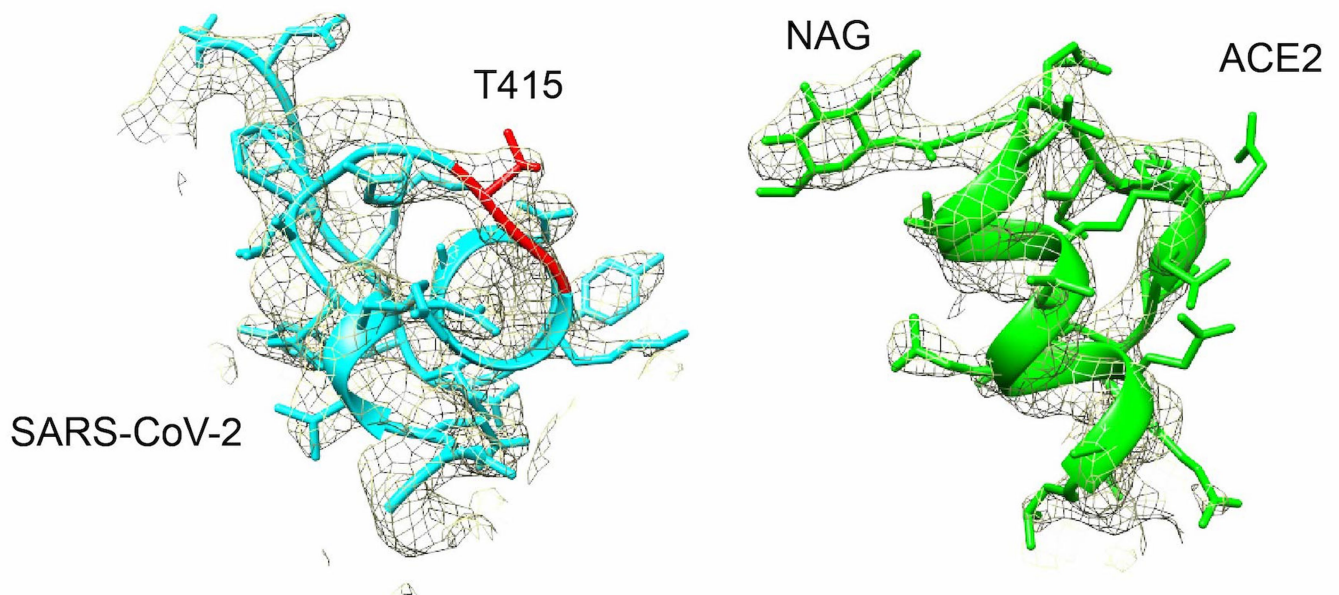


Extended Data Fig. 3 | Structural comparisons of the RBDs of SARS-CoV-2 and SARS-CoV and their binding modes to the ACE2 receptor. a, Alignment of the SARS-CoV-2 RBD (core in cyan and RBM in red) and SARS-CoV RBD (core in orange and RBM in blue) structures. **b,** Structural alignment of the

SARS-CoV-2 RBD-ACE2 and SARS-CoV RBD-ACE2 complexes. The SARS-CoV-2 RBD is shown in cyan and red, its interacting ACE2 is shown in green. The SARS-CoV RBD is shown in orange and blue, its interacting ACE2 is shown in salmon. The PDB code for the SARS-CoV RBD-ACE2 complex is 2AJF.



Extended Data Fig. 4 | Surface plasmon resonance sensorgrams. Binding curves of immobilized human ACE2 with the SARS-CoV-2 RBD (left) and SARS-CoV RBD (right). Data are shown as different coloured lines and the best fit of the data to a 1:1 binding model is shown in black.

a**b**

Extended Data Fig. 5 | Asn90-linked glycans of ACE2. **a**, The interface between Asn90-linked glycans of ACE2 (salmon) and SARS-CoV RBD (orange). **b**, The interface between Asn90-linked glycan of ACE2 (green) and SARS-CoV-2 RBD (cyan). The $2F_o - F_c$ electron densities contoured at 1.5σ are also shown.

	SARS-CoV RBD-ACE2
Data collection	
Space group	P4 ₁ 2 ₁ 2
Cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	104.67, 104.67, 228.72
α , β , γ , (°)	90, 90, 90
Resolution (Å)	53.1-2.45(2.54-2.45) *
<i>R</i> _{sym} or <i>R</i> _{merge}	0.118(2.70)
<i>I</i> / <i>sI</i>	24.2(1.7)
Completeness (%)	99.90(99.98)
Redundancy	26.1(27.3)
Refinement	
Resolution (Å)	53.1-2.45
No. reflections	47555
<i>R</i> _{work} / <i>R</i> _{free}	19.6/23.7
No. atoms	
Protein	6419
Ligand/ion	57
Water	80
<i>B</i> -factors	
Protein	64.7
Ligand/ion	91.4
Water	58.1
R.m.s. deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.21

One crystal was used.

*Values in parentheses are for the highest-resolution shell.

SARS-CoV-2	ACE2	SARS-CoV	ACE2
RBD		RBD	
K417	Q24	R426	Q24
G446	T27	Y436	T27
Y449	F28	Y440	F28
Y453	D30	Y442	K31
L455	K31	L443	H34
F456	H34	L472	E37
A475	E35	N473	D38
F486	E37	Y475	Y41
N487	D38	N479	Q42
Y489	Y41	G482	L45
Q493	Q42	Y484	L79
G496	L79	T486	M82
Q498	M82	T487	Y83
T500	Y83	G488	Q325
N501	N330	I489	E329
G502	K353	Y491	N330
Y505	G354		K353
	D355		G354
	R357		D355
	R393		R357

A distance cut-off of 4 Å was used.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

NCBI was used for downloading the published SARS-CoV-2 sequences to do the research

Data analysis

Aquarium pipeline, CCP4 v7.0, COOT 0.8.6 and PHENIX 1.15.2-3472 were used at the determination of complex structure for data processing, model building and refinement. PyMOL 1.8.x was used to generate the structural figures.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Crystal structure presented in this work has been deposited in the Protein Data Bank (PDB) and are available with accession codes 6M0J.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Two separate proteins were purified and formed a complex.
Data exclusions	No data excluded.
Replication	The crystal of the complex were obtained in duplicate. All attempts at replication were successful.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	SF9 and Hi5 cells were bought from ATCC
Authentication	SF9 : refered to website: https://www.atcc.org/products/all/CRL-1711.aspx Hi5 : refered to website These two cell lines are all available in commercial company.
Mycoplasma contamination	We confirm that all cell lines were negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used

Structural basis of receptor recognition by SARS-CoV-2

<https://doi.org/10.1038/s41586-020-2179-y>

Received: 16 February 2020

Accepted: 20 March 2020

Published online: 30 March 2020

 Check for updates

Jian Shang^{1,3}, Gang Ye^{1,3}, Ke Shi^{2,3}, Yushun Wan^{1,3}, Chuming Luo¹, Hideki Aihara², Qibin Geng¹, Ashley Auerbach¹ & Fang Li^{1✉}

A novel severe acute respiratory syndrome (SARS)-like coronavirus (SARS-CoV-2) recently emerged and is rapidly spreading in humans, causing COVID-19^{1,2}. A key to tackling this pandemic is to understand the receptor recognition mechanism of the virus, which regulates its infectivity, pathogenesis and host range. SARS-CoV-2 and SARS-CoV recognize the same receptor—angiotensin-converting enzyme 2 (ACE2)—in humans^{3,4}. Here we determined the crystal structure of the receptor-binding domain (RBD) of the spike protein of SARS-CoV-2 (engineered to facilitate crystallization) in complex with ACE2. In comparison with the SARS-CoV RBD, an ACE2-binding ridge in SARS-CoV-2 RBD has a more compact conformation; moreover, several residue changes in the SARS-CoV-2 RBD stabilize two virus-binding hotspots at the RBD–ACE2 interface. These structural features of SARS-CoV-2 RBD increase its ACE2-binding affinity. Additionally, we show that RaTG13, a bat coronavirus that is closely related to SARS-CoV-2, also uses human ACE2 as its receptor. The differences among SARS-CoV-2, SARS-CoV and RaTG13 in ACE2 recognition shed light on the potential animal-to-human transmission of SARS-CoV-2. This study provides guidance for intervention strategies that target receptor recognition by SARS-CoV-2.

The sudden emergence and rapid spread of SARS-CoV-2 is endangering global health and economy^{1,2}. SARS-CoV-2 has caused many more infections, deaths and economic disruptions than SARS-CoV in 2002–2003^{5,6}. The origin of SARS-CoV-2 remains unclear. Bats are considered the original source of SARS-CoV-2 because a closely related coronavirus, RaTG13, has been isolated from bats⁷. However, the molecular events that led to the possible bat-to-human transmission of SARS-CoV-2 are unknown. Clinically approved vaccines or drugs that specifically target SARS-CoV-2 are also lacking. Receptor recognition by coronaviruses is an important determinant of viral infectivity, pathogenesis and host range^{8,9}. It presents a major target for vaccination and antiviral strategies¹⁰. Here we elucidate the structural and biochemical mechanisms of receptor recognition by SARS-CoV-2.

Receptor recognition by SARS-CoV has been extensively studied. A virus-surface spike protein mediates the entry of coronavirus into host cells. The spike protein of SARS-CoV contains a RBD that specifically recognizes ACE2 as its receptor^{3,4}. A series of crystal structures of the SARS-CoV RBD from different strains in complex with ACE2 from different hosts has previously been determined^{3,11,12}. These structures showed that SARS-CoV RBD contains a core and a receptor-binding motif (RBM); the RBM mediates contacts with ACE2. The surface of ACE2 contains two virus-binding hotspots that are essential for SARS-CoV binding. Several naturally selected mutations in the SARS-CoV RBM surround these hotspots and regulate the infectivity, pathogenesis, and cross-species and human-to-human transmissions of SARS-CoV^{3,11,12}.

Because of the sequence similarity between the spike proteins of SARS-CoV and SARS-CoV-2, it was recently predicted that SARS-CoV-2

also uses ACE2 as its receptor¹³, which has been validated by other studies^{7,14–16}. Here we determined the structural basis of receptor recognition by SARS-CoV-2 and compared the ACE2-binding affinity among SARS-CoV-2, SARS-CoV and RaTG13. Our findings identify the molecular and structural features of the SARS-CoV-2 RBM that result in tight ACE2 binding. They provide insights into the animal origin of SARS-CoV-2, and can help to guide intervention strategies that target SARS-CoV-2–ACE2 interactions.

To understand the structural basis of ACE2 recognition by SARS-CoV-2, we aimed to crystallize the SARS-CoV-2 RBD–ACE2 complex. Our strategy was informed by previous crystallization of the SARS-CoV RBD–ACE2 complex³. In this crystal form, the core of the SARS-CoV RBD (along with the ACE2 surface) was mainly involved in crystal lattice contact; the essential ACE2-binding residues in the SARS-CoV RBM were buried at the RBD–ACE2 interface and did not affect crystallization. To facilitate crystallization, we designed a chimeric RBD that uses the core from the SARS-CoV RBD as the crystallization scaffold and the RBM from SARS-CoV-2 as the functionally relevant unit (Fig. 1a and Extended Data Fig. 1). To further enhance crystallization, we improved the ACE2-binding affinity of the chimeric RBD by keeping a short loop from the SARS-CoV RBM, which maintains a strong salt bridge between Arg426 of the RBD and Glu329 of ACE2 (Extended Data Fig. 2a). This loop sits on the side of the binding interface, away from the main binding interface. We expressed and purified the chimeric RBD and ACE2, and crystallized the complex under the same conditions and in the same crystal form as those used for the SARS-CoV RBD–ACE2 complex. On the basis of X-ray diffraction data,

¹Department of Veterinary and Biomedical Sciences, University of Minnesota, Saint Paul, MN, USA. ²Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, USA. ³These authors contributed equally: Jian Shang, Gang Ye, Ke Shi, Yushun Wan. ✉e-mail: lifang@umn.edu

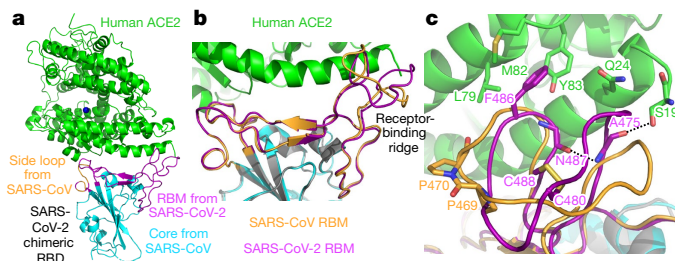


Fig. 1 | Structure of the SARS-CoV-2 chimeric RBD complexed with ACE2. **a**, Crystal structure of the SARS-CoV-2 chimeric RBD complexed with ACE2. ACE2 is shown in green. The RBD core is shown in cyan. The RBM is shown in magenta. A side loop in RBM is shown in orange. A zinc ion in ACE2 is shown in blue. **b**, Comparison of the conformations of the ridge in SARS-CoV-2 RBM (magenta) and SARS-CoV RBM (orange). **c**, Comparison of the conformations of the ridge from another viewing angle. In the SARS-CoV RBM, a proline-proline-alanine motif is shown. In the SARS-CoV-2 RBM, a newly formed hydrogen bond, Phe486, and some of the interactions of the ridge with the N-terminal helix of ACE2 are shown.

we determined the structure of the chimeric RBD–ACE2 complex by molecular replacement using the structure of the SARS-CoV RBD–ACE2 complex as the search template. We refined the structure to 2.68 Å (Extended Data Table 1 and Extended Data Fig. 3). The structure of this chimeric RBD–ACE2 complex, particularly in the RBM region, is highly similar to another recently determined structure of the SARS-CoV-2 wild-type RBD–ACE2 complex¹⁷, confirming that the chimeric RBD is a successful design.

The overall structure of the chimeric RBD–ACE2 complex is similar to that of the SARS-CoV RBD–ACE2 complex (Fig. 1a). Similar to the SARS-CoV RBM, SARS-CoV-2 RBM forms a gently concave surface with a ridge on one side; it binds to the exposed outer surface of the claw-like structure of ACE2 (Fig. 1a). The strong salt bridge between SARS-CoV RBD and ACE2 became a weaker (as judged by the longer distance of the interaction), but still energetically favourable, N–O bridge between Arg439 of the chimeric RBD and Glu329 of ACE2¹⁸ (Extended Data Fig. 2b). In comparison to the SARS-CoV RBM, the SARS-CoV-2 RBM forms a larger binding interface and more contacts with ACE2 (Extended Data Fig. 4a, b). Our structural model also contained glycans attached to four ACE2 sites and one RBD site (Extended Data Fig. 5a). The glycan attached to Asn90 of ACE2 forms a hydrogen bond with Arg408 of the RBD core (Extended Data Fig. 5b); this glycan-interacting arginine is conserved between SARS-CoV-2 and SARS-CoV (Extended Data Fig. 1). The overall structural similarity in ACE2 binding by SARS-CoV-2 and SARS-CoV supports a close evolutionary relationship between the two viruses.

We measured the binding affinities between each of the three RBMs (SARS-CoV-2, chimeric and SARS-CoV) and ACE2 using surface plasmon resonance (SPR) (Extended Data Figs. 4c, 6). We found that the chimeric RBD has a higher ACE2-binding affinity than the SARS-CoV-2 RBD, consistent with the introduced N–O bridge between the chimeric RBD and ACE2. Both the chimeric and SARS-CoV-2 RBMs have significantly higher ACE2-binding affinities than the SARS-CoV RBD. These dissociation constant K_d values are consistent with other SPR studies^{12,19}, although the exact K_d values vary depending on the specific approaches of each SPR experiment (Extended Data Table 2). Here we investigate the structural differences between the RBMs of SARS-CoV-2 and SARS-CoV that account for their different ACE2-binding affinities.

A marked structural difference between the RBMs of SARS-CoV-2 and SARS-CoV is the conformation of the loops in the ACE2-binding ridge (Fig. 1b, c). In both RBMs, one of the ridge loops contains an essential disulfide bond and the region between the disulfide-bond-forming cysteines is variable (Fig. 1c and Extended Data Fig. 1). Specifically, human and civet SARS-CoV strains and bat coronavirus Rs3367 all contain a three-residue motif proline-proline-alanine in this loop;

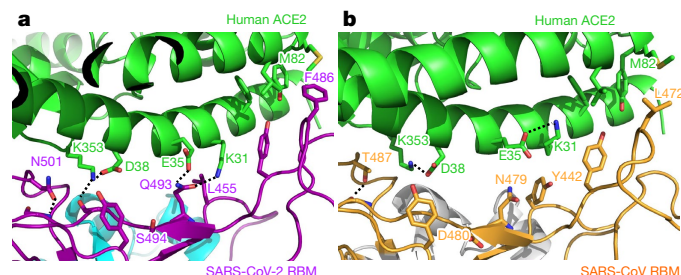


Fig. 2 | Structural details at the interface between the SARS-CoV-2 RBM and ACE2. **a**, The interface between the SARS-CoV-2 RBM and ACE2. **b**, The interface between SARS-CoV RBM and ACE2.

the tandem prolines allow the loop to take a sharp turn. By contrast, SARS-CoV-2 and bat coronavirus RaTG13 both contain a four-residue motif glycine-valine/glutamine-glutamate/threonine-glycine; the two relatively bulky residues and two flexible glycines enable the loop to take a different conformation (Fig. 1c and Extended Data Fig. 1). Because of these structural differences, an additional main-chain hydrogen bond forms between Asn487 and Ala475 in the SARS-CoV-2 RBM, causing the ridge to take a more compact conformation and the loop containing Ala475 to move closer to ACE2 (Fig. 1c). As a consequence, the ridge in the SARS-CoV-2 RBM forms more contacts with the N-terminal helix of ACE2 (Extended Data Fig. 4b). For example, the N-terminal residue Ser19 of ACE2 forms a new hydrogen bond with the main chain of Ala475 of the SARS-CoV-2 RBM, and Gln24 in the N-terminal helix of ACE2 also forms a new contact with the SARS-CoV-2 RBM (Fig. 1c and Extended Data Fig. 4b). Moreover, compared with the corresponding Leu472 of the SARS-CoV RBM, Phe486 of the SARS-CoV-2 RBM points in a different direction and inserts into a hydrophobic pocket involving Met82, Leu79 and Tyr83 of ACE2 (Figs. 1c, 2a, b). In comparison to the SARS-CoV RBM, these structural changes in the SARS-CoV-2 RBM are more favourable for ACE2 binding.

In comparison to the SARS-CoV RBM–ACE2 interface, subtle yet functionally important structural changes take place near the two virus-binding hotspots at the SARS-CoV-2 RBM–ACE2 interface (Fig. 2a, b). At the SARS-CoV–ACE2 interface, two virus-binding hotspots were previously identified^{11,12}: hotspot Lys31 (that is, hotspot 31) consists of a salt bridge between Lys31 and Glu35, and hotspot Lys353 (that is, hotspot 353) consists of a salt bridge between Lys353 and Asp38. Both salt bridges are weak, as judged by the relatively long distance of these interactions. Burial of these weak salt bridges in hydrophobic environments on virus binding would enhance their energy, owing to a reduction in the dielectric constant. This process is facilitated by interactions between the hotspots and nearby RBD residues. First, at the SARS-CoV RBM–ACE2 interface, hotspot 31 requires support from Tyr442 of the SARS-CoV RBM (Fig. 2b). In comparison, at the SARS-CoV-2 RBM–ACE2 interface, Leu455 of the SARS-CoV-2 RBM (corresponding to Tyr442 of the SARS-CoV RBM) has a less bulky side chain, providing less support to Lys31 of ACE2. As a result, the structure of hotspot 31 has rearranged: the salt bridge between Lys31 and Glu35 breaks apart, and each of the residues forms a hydrogen bond with Gln493 of the SARS-CoV-2 RBM (Fig. 2a). Second, at the SARS-CoV RBM–ACE2 interface, hotspot 353 requires support from the side-chain methyl group of Thr487 of the SARS-CoV RBM, whereas the side-chain hydroxyl group of Thr487 forms a hydrogen bond with the RBM main chain (which fixes the conformation of the Thr487 side chain) (Fig. 2b). In comparison, at the SARS-CoV-2 RBM–ACE2 interface, Asn501 of the SARS-CoV-2 RBM also has its conformation fixed through a hydrogen bond between its side chain and the RBM main chain; correspondingly, its side chain provides less support to hotspot 353 than the corresponding Thr487 of the SARS-CoV RBM does (Fig. 2a). Consequently, Lys353

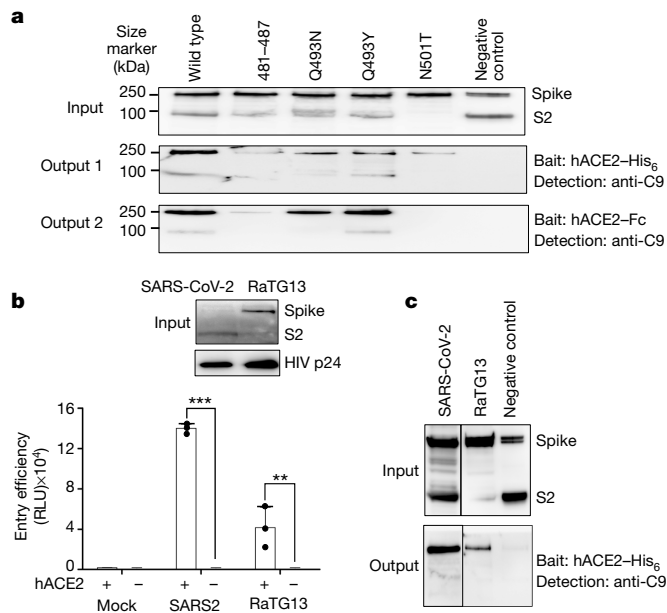


Fig. 3 | Biochemical data showing the interactions between SARS-CoV-2 or bat RaTG13 spike and ACE2. a, Protein pull-down assay using ACE2 as the bait and cell-associated SARS-CoV-2 spike molecules (wild type and mutants) as the targets. Top, cell-expressed SARS-CoV-2 spike. Middle, pull-down results using His₆-tagged ACE2. Bottom, pull-down results using Fc-tagged ACE2. MERS-CoV spike was used as a negative control. **b**, Entry of SARS-CoV-2 and bat RaTG13 pseudoviruses into ACE2-expressing cells. Top, packaged SARS-CoV-2 and bat RaTG13 pseudoviruses. HIV p24 was detected as an internal control. Bottom, pseudovirus entry efficiency. Mock, no pseudoviruses. Data are mean \pm s.d. A comparison (two-tailed Student's *t*-test) between SARS-CoV-2 with ACE2 ($n = 3$ independent samples) and SARS-CoV-2 without ACE2 ($n = 4$ independent samples) showed a significant difference ($P < 1.16 \times 10^{-8}$). A comparison (two-tailed Student's *t*-test) between RaTG13 with ACE2 ($n = 3$ independent samples) and RaTG13 without ACE2 ($n = 4$ independent samples) showed a significant difference, $P = 0.0097$. Individual data points are shown as black dots. *** $P < 0.001$; ** $P < 0.01$. **c**, Protein pull-down assay using ACE2 as the bait and cell-associated RaTG13 spike as the target. All experiments were repeated independently three times with similar results.

of ACE2 takes a slightly different conformation, forming a hydrogen bond with the main chain of the SARS-CoV-2 RBM while maintaining the salt bridge with Asp38 of ACE2 (Fig. 2a). Thus, both hotspots have adjusted to the reduced support from nearby RBD residues, yet still become well-stabilized at the SARS-CoV-2 RBM–ACE2 interface.

To corroborate the structural observations, we characterized ACE2-binding affinities of the SARS-CoV-2 spike that contains mutations in critical ACE2-interacting residues. To this end, protein pull-down assays were performed, with purified recombinant ACE2 as the bait and cell-associated SARS-CoV-2 spike as the target (Fig. 3a). For cross-validation, we used ACE2 with two different tags, His₆ and Fc. The SARS-CoV-2 spike contained one of the following RBM changes: 481–487 (481-NGVEGFN-487 in SARS-CoV-2 were mutated to TPPALN as in SARS-CoV), Q493N (Gln493 in SARS-CoV-2 was mutated to an asparagine as in human SARS-CoV), Q493Y (Gln493 in SARS-CoV-2 was mutated to a tyrosine as in bat RaTG13) and N501T (Asn501 in SARS-CoV-2 was mutated to a threonine as in human SARS-CoV). The results showed that all of these introduced mutations reduced the ACE2-binding affinity of the SARS-CoV-2 spike. They confirm that the structural features of the SARS-CoV-2 RBM, including the ACE2-binding ridge and the hotspots-stabilizing residues, all contribute to the high ACE2-binding affinity of SARS-CoV-2.

Having compared ACE2 recognition by SARS-CoV-2 and SARS-CoV, we further investigated human ACE2 binding by bat RaTG13. To this end,

we performed a pseudovirus entry assay in which retroviruses pseudotyped with RaTG13 spike (that is, RaTG13 pseudoviruses) were used to enter ACE2-expressing human cells (Fig. 3b). The results showed that RaTG13 pseudovirus entry into the cells depends on ACE2. Additionally, RaTG13 spike was not cleaved on the pseudovirus surface. SARS-CoV-2 pseudovirus entry also depends on ACE2, but its spike was cleaved to S2 on the pseudovirus surface (probably because of a furin site insertion¹⁶) (Fig. 3b). Moreover, we performed a protein pull-down assay using ACE2 as the bait and cell-associated RaTG13 spike as the target (Fig. 3c). We found that the RaTG13 spike was pulled down by ACE2. Therefore, similar to SARS-CoV-2, bat RaTG13 binds to human ACE2 and can use human ACE2 as its entry receptor.

The current SARS-CoV-2 outbreak has become a global pandemic. Previous structural studies on SARS-CoV have established receptor recognition as an important determinant of SARS-CoV infectivity, pathogenesis and host range⁹. On the basis of the structural information presented here, along with biochemical data, we discuss the receptor recognition and evolution of SARS-CoV-2.

We will first discuss how well SARS-CoV-2 recognizes ACE2 in comparison to SARS-CoV. We show that, compared with SARS-CoV, the SARS-CoV-2 RBM contains structural changes in the ACE2-binding ridge, largely caused by a four-residue motif (residues 482–485: Gly–Val–Glu–Gly). This structural change allows the ridge to become more compact and form better contacts with the N-terminal helix of ACE2 (Fig. 1b, c). In addition, Phe486 of the SARS-CoV-2 RBM inserts into a hydrophobic pocket (Fig. 1c). The corresponding residue in the SARS-CoV RBM is a leucine, which probably forms a weaker contact with ACE2 owing to its smaller side chain. Finally, both virus-binding hotspots are more stabilized at the RBM–ACE2 interface through interactions with the SARS-CoV-2 RBM. As previous studies have shown^{11,12}, these hotspots on ACE2 are important for coronavirus binding, because they involve two lysine residues that need to be accommodated properly in hydrophobic environments. Neutralizing the charges of the lysines is key to the binding of coronavirus RBDs to ACE2. The SARS-CoV-2 RBM has evolved strategies to stabilize the two hotspots: Gln493 and Leu455 stabilize hotspot 31, whereas Asn501 stabilizes hotspot 353 (Fig. 2a). Our biochemical data confirm that the SARS-CoV-2 RBD has a significantly higher ACE2-binding affinity than the SARS-CoV RBD and that the above structural features of the SARS-CoV-2 RBM contribute to the high ACE2-binding affinity of SARS-CoV-2 RBD (Fig. 3a). Thus, both structural and biochemical data reveal that the SARS-CoV-2 RBD recognizes ACE2 better than SARS-CoV RBD does.

Next, we investigated how SARS-CoV-2 may have been transmitted from bats to humans. First, we found that bat RaTG13 uses human ACE2 as its receptor (Fig. 3b, c), suggesting that RaTG13 may infect humans. Second, as with SARS-CoV-2, bat RaTG13 RBM contains a similar four-residue motif in the ACE2-binding ridge, supporting the notion that SARS-CoV-2 may have evolved from RaTG13 or a RaTG13-related bat coronavirus (Extended Data Table 3 and Extended Data Fig. 7). Third, the L486F, Y493Q and D501N residue changes from RaTG13 to SARS-CoV-2 enhance ACE2 recognition and may have facilitated the bat-to-human transmission of SARS-CoV-2 (Extended Data Table 3 and Extended Data Fig. 7). A lysine-to-asparagine mutation at the 479 position in the SARS-CoV RBD (corresponding to the 493 position in the SARS-CoV-2 RBD) enabled SARS-CoV to infect humans³. Fourth, Leu455 contributes favourably to ACE2 recognition, and it is conserved between RaTG13 and SARS-CoV-2; its presence in the SARS-CoV-2 RBM may be important for the bat-to-human transmission of SARS-CoV-2 (Extended Data Table 3 and Extended Data Fig. 7). Host and viral factors other than receptor recognition also have important roles in the cross-species transmission of coronaviruses^{20,21}. Nevertheless, the identified receptor-binding features of the SARS-CoV-2 RBM may have facilitated SARS-CoV-2 to transmit from bats to humans (Extended Data Fig. 7).

We then examined whether intermediate hosts were involved in the potential bat-to-human transmission of SARS-CoV-2. Because bat coronavirus RaTG13 binds to human ACE2, one possibility is that there is no intermediate host. Alternatively, pangolins have been proposed to be an intermediate host²². The structural information provided in this study enables us to inspect and understand the important RBM residues in coronaviruses isolated from pangolins. Two coronaviruses, CoV-pangolin/GD and CoV-pangolin/GX, have been isolated from pangolins from two different locations in China: Guangdong (GD) and Guangxi (GX), respectively. The RBM of the CoV-pangolin/GD contains Leu455, the 482–485 loop, Phe486, Gln493 and Asn501 (Extended Data Table 3), all of which are favourable for ACE2 recognition. The RBM of CoV-pangolin/GX contains Leu455 and the 482–485 loop, both of which are favourable for ACE2 recognition, and it also contains Leu486, Glu493 and Thr501 (Extended Data Table 3), all of which are less favourable for ACE2 recognition. Therefore, CoV-pangolin/GD potentially recognizes human ACE2 well, whereas CoV-pangolin/GX does not. Hence, pangolins from Guangdong, but not pangolins from Guangxi, could potentially pass coronaviruses to humans. However, many other factors determine the cross-species transmission of coronaviruses^{20,21}, and the above analysis will need to be verified experimentally.

Finally, this study helps to inform intervention strategies. First, neutralizing monoclonal antibodies that target the SARS-CoV-2 RBM can prevent the virus from binding to ACE2, and are therefore promising antiviral drugs. Our structure has laid out all of the functionally important epitopes in the SARS-CoV-2 RBM that can potentially be targeted by neutralizing antibody drugs. Thus, this study can help to guide the development and optimization of these antibody drugs. Second, the RBD itself can function as a subunit vaccine^{10,23}. The functionally important epitopes in the SARS-CoV-2 RBM that were identified in this study can guide structure-based design of highly efficacious RBD vaccines. Such a structure-based strategy for subunit vaccine design has previously been developed²⁴. This strategy may be helpful in designing SARS-CoV-2 RBD vaccines. Overall, this study can help to inform structure-based intervention strategies that target receptor recognition by SARS-CoV-2.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2179-y>.

- Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
- Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
- Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864–1868 (2005).
- Li, W. et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454 (2003).
- Lee, N. et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. *N. Engl. J. Med.* **348**, 1986–1994 (2003).
- Peiris, J. S. M. et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325 (2003).
- Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
- Perlman, S. & Netland, J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat. Rev. Microbiol.* **7**, 439–450 (2009).
- Li, F. Structure, function, and evolution of coronavirus spike proteins. *Annu. Rev. Virol.* **3**, 237–261 (2016).
- Du, L. et al. The spike protein of SARS-CoV — a target for vaccine and therapeutic development. *Nat. Rev. Microbiol.* **7**, 226–236 (2009).
- Li, F. Structural analysis of major species barriers between humans and palm civets for severe acute respiratory syndrome coronavirus infections. *J. Virol.* **82**, 6984–6991 (2008).
- Wu, K., Peng, G., Wilken, M., Geraghty, R. J. & Li, F. Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. *J. Biol. Chem.* **287**, 8904–8911 (2012).
- Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* **94**, e00127–20 (2020).
- Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).
- Hoffmann, M. et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* <https://doi.org/10.1016/j.cell.2020.02.052> (2020).
- Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* <https://doi.org/10.1016/j.cell.2020.02.058> (2020).
- Lan, J. et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* <https://doi.org/10.1038/s41586-020-2180-5> (2020).
- Pylaeva, S., Brehm, M. & Sebastiani, D. Salt bridge in aqueous solution: strong structural motifs but weak enthalpic effect. *Sci. Rep.* **8**, 13626 (2018).
- Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
- Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
- Yang, Y. et al. Receptor usage and cell entry of bat coronavirus HKU4 provide insight into bat-to-human transmission of MERS coronavirus. *Proc. Natl Acad. Sci. USA* **111**, 12516–12521 (2014).
- Lam, T. T. et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature* <https://doi.org/10.1038/s41586-020-2169-0> (2020).
- Du, L. et al. MERS-CoV spike protein: a key target for antivirals. *Expert Opin. Ther. Targets* **21**, 131–143 (2017).
- Du, L. et al. Introduction of neutralizing immunogenicity index to the rational design of MERS coronavirus subunit vaccines. *Nat. Commun.* **7**, 13473 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Plasmids

SARS-CoV-2 spike (GenBank accession number QHD43416.1), SARS-CoV spike (GenBank accession number AFR58740.1), RaTG13 spike (GenBank accession number QHR63300.2) and ACE2 (GenBank accession number NM_021804) were all synthesized (GenScript Biotech). SARS-CoV-2, SARS-CoV, chimeric RBDs (see Extended Data Fig. 1 for residue ranges of RBDs) and ACE2 ectodomain (residues 1–615) were subcloned into pFastBac vector (Life Technologies) with a N-terminal honeybee melittin signal peptide and a C-terminal His₆-tag. The ACE2 ectodomain (residues 1–615) with a C-terminal Fc-tag was also constructed.

Protein expression and purification

All of the proteins were prepared from Sf9 insect cells using the Bac-to-Bac system (Life Technologies) as previously described³. In brief, the His₆-tagged proteins were collected from cell culture medium, purified using a Ni-NTA column, purified further using a Superdex200 gel filtration column (GE Healthcare) and stored in a buffer containing 20 mM Tris pH 7.2 and 200 mM NaCl. The Fc-tagged protein was purified in the same way as the His₆-tagged proteins, except that the protein A column replaced the Ni-NTA column in the procedure.

Crystallization and structure determination

To purify the RBD–ACE2 complex, ACE2 and RBD were incubated together, and the complex was purified using Superdex200 gel-filtration chromatography. RBD–ACE2 crystals were grown in sitting drops at room temperature over wells containing 100 mM Tris (pH 8.5), 18–20% PEG 6000 and 100 mM NaCl. Crystals were soaked briefly in 100 mM Tris (pH 8.5), 30% PEG 6000, 100 mM NaCl and 30% ethylene glycol before being flash-frozen in liquid nitrogen. X-ray diffraction data were collected at the Advanced Photon Source beamline 24-ID-E. The structure was determined by molecular replacement using the structure of SARS-CoV RBD complexed with ACE2 as the search template (Protein Data Bank (PDB) 2AJF). Structure data and refinement statistics are shown in Extended Data Table 1.

Protein–protein binding assay

The SPR assays using a Biacore 2000 system (GE Healthcare) were carried out as described previously¹². In brief, different RBDs were covalently immobilized to a CM5 sensor chip through their amine groups (GE Healthcare). The running buffer contained 10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA and 0.05% Tween-20. Serial dilutions of purified recombinant ACE2 were injected ranging in concentration from 5 to 80 nM for the SARS-CoV-2 RBD and chimeric RBD, and from 20 to 320 nM for the SARS-CoV RBD. The resulting data were fit to a 1:1 binding model using Biacore Evaluation Software (GE Healthcare).

The protein pull-down assay was performed using a Dynabeads His-Tag Isolation and Pull-down kit (Invitrogen) and a Dynabeads Protein A for Immunoprecipitation kit (Invitrogen) according to the manufacturers' manual. In brief, 150 µl indicated Dynabeads were washed with phosphate-buffered saline (PBS) and incubated with either 5 µg ACE2–His₆ (ACE2 with a C-terminal His₆-tag) or 5 µg ACE2–Fc (ACE2 with a C-terminal Fc-tag) on a roller at room temperature for 30 min. After incubation, ACE2-bound beads were washed three times with 1 ml PBST buffer (PBS and 0.05% Tween-20) on a roller for 10 min and then were aliquoted into different tubes for use. To prepare the cell-associated coronavirus spike protein, HEK293T cells were transfected with a pcDNA3.1(+) plasmid encoding coronavirus spike (containing a C-terminal C9-tag); 48 h after transfection, the spike-expressing cells

were lysed using a sonicator in immunoprecipitation assay buffer (20 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA and 1% Triton X-100, supplemented with protease inhibitors) and centrifuged at 12,000g for 2 min. The supernatants (containing solubilized SARS-CoV-2 spike) were transferred to mix with the ACE2-bound beads in 2-ml tubes separately (spike was in excess of ACE2). After a 1-h incubation on a roller at room temperature, beads were washed three times with PBST buffer and the bound proteins were eluted using elution buffer (300 mM imidazole, 50 mM sodium phosphate pH 8.0, 300 mM NaCl, 0.01% Tween-20 for ACE2–His₆-bound beads; 0.1 M citric acid pH 2.7 for ACE2–Fc-bound beads). The samples were then subjected to SDS–PAGE and analysed by western blotting using an anti-C9 tag antibody.

Coronavirus-spike-mediated pseudovirus entry assay

The pseudovirus entry assay was performed as described previously²¹. In brief, HEK293T cells were co-transfected with a luciferase-expressing HIV-1 genome plasmid (pNL4-3.luc.RE) and a plasmid encoding SARS-CoV-2 spike or RaTG13 spike. Pseudoviruses were collected 72 h after transfection, and were used to enter recipient cells (HEK293T cells exogenously expressing ACE2). After incubation of pseudoviruses with recipient cells at 37 °C for 6 h, the medium was changed and cells were incubated for an additional 60 h. Cells were then washed with PBS buffer and lysed. Aliquots of cell lysates were transferred to Optiplate-96 (PerkinElmer), followed by the addition of luciferase substrate. Relative light units were measured using an EnSpire plate reader (PerkinElmer). All measurements were carried out on at least three independent biological samples.

Analyses of protein contact residues and protein buried surface areas

Protein contact residues were analysed using the LigPlot⁺ program (v.1.4.5) (<https://www.ebi.ac.uk/thornton-srv/software/LigPlus/>). Protein buried surface areas were analysed using PDBePISA tool (<http://pdbe.org/pisa/>).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Coordinates and structure factors have been deposited to the Protein Data Bank with accession number 6VW1.

- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
- Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Sui, J. et al. Potent neutralization of severe acute respiratory syndrome (SARS) coronavirus by a human mAb to S1 protein that blocks receptor association. *Proc. Natl Acad. Sci. USA* **101**, 2536–2541 (2004).
- Li, W. et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.* **24**, 1634–1643 (2005).
- Sun, C. et al. SARS-CoV-2 and SARS-CoV spike-RBD structure and receptor binding comparison and potential implications on neutralizing antibody and vaccine development. Preprint at <https://doi.org/10.1101/2020.02.16.951723> (2020).
- Pesce, A. J. & Michael, J. G. Artifacts and limitations of enzyme immunoassay. *J. Immunol. Methods* **150**, 111–119 (1992).

Acknowledgements This work was supported by NIH grants R01AI089728 and R01AI110700 (to F.L.) and R35GM118047 (to H.A.) and is based on research conducted at the Northeastern Collaborative Access Team beamlines, which are supported by NIH grants P30GM124165 and S10OD021527, and DOE contract DE-AC02-06CH11357. We thank staff at Advanced Photon Source beamline 24-ID-E for assistance in data collection and Y. V. Jiang for statistical consultation and edits to the manuscript.

Author contributions J.S. conceptualized the project, expressed and purified proteins, performed crystallization, carried out protein pull-down experiments and the pseudovirus entry assay, and reviewed the manuscript. G.Y. performed crystallization, determined and refined the structure, analysed the structure, performed the SPR experiment, and reviewed the manuscript.

Article

K.S. collected X-ray diffraction data, determined and refined the structure, analysed the structure, and reviewed the manuscript. Y.W. conceptualized the project, expressed and purified proteins, performed protein pull-down experiments and the pseudovirus entry assay, and reviewed the manuscript. C.L. performed protein pull-down experiments and the pseudovirus entry assay, and reviewed the manuscript. H.A. provided resources, analysed the structure, and reviewed the manuscript. Q.G. performed protein pull-down experiments and the pseudovirus entry assay, and reviewed the manuscript. A.A. expressed and purified proteins, and reviewed the manuscript. F.L. conceptualized and supervised the project, provided resources, guided the experiments and data analysis, and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2179-y>.

Correspondence and requests for materials should be addressed to F.L.

Peer review information *Nature* thanks Lijun Rong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.


```

SARS-human      306 RVVPS GDVVRFPNIT NLCPFGEVFN ATKFPSVYAW ERKRISNCVA DYSLVYNSTF 360
SARS-civet      306 RVVPS GDVVRFPNIT NLCPFGEVFN ATKFPSVYAW ERKRISNCVA DYSLVYNSTS 360
CoV-pangolin/GD 319 RVQPT ESIVRFPNIT NLCPFGEVFN ATTFASVYAW NRKRISNCVA DYSLVYNSTS 373
CoV-pangolin/GX 319 RVQPT ISIVRFPNIT NLCPFGEVFN ASKFASVYAW NRKRISNCVA DYSLVYNSTS 373
Rs3367-bat      307 RVAPS KEVVRFPNIT NLCPFGEVFN ATTFPSVYAW ERKRISNCVA DYSLVYNSTS 361
RaTG13-bat      319 RVQPT DSIVRFPNIT NLCPFGEVFN ATTFASVYAW NRKRISNCVA DYSLVYNSTS 373
SARS-CoV-2      319 RVQPT ESIVRFPNIT NLCPFGEVFN ATRFASVYAW NRKRISNCVA DYSLVYNSTAS 373
                **.*: ..:***** **:****** *:.*:***** :*:***** *****:

SARS-human      FSTFKCYGVS ATKLNLCFS NVYADSFVVK GDDVQIAPG QTGVIADYNY KLPDDFMGCV 420
SARS-civet      FSTFKCYGVS ATKLNLCFS NVYADSFVVK GDDVQIAPG QTGVIADYNY KLPDDFMGCV 420
CoV-pangolin/GD FSTFKCYGVS PTKLNLCFT NVYADSFVVR GDEVQIAPG QTGRIADYNY KLPDDFTGCV 433
CoV-pangolin/GX FSTFKCYGVS PTKLNLCFT NVYADSFVVK GDEVQIAPG QTGVIADYNY KLPDDFTGCV 433
Rs3367-bat      FSTFKCYGVS ATKLNLCFS NVYADSFVVK GDDVQIAPG QTGVIADYNY KLPDDFTGCV 421
RaTG13-bat      FSTFKCYGVS PTKLNLCFT NVYADSFVIT GDEVQIAPG QTGVIADYNY KLPDDFTGCV 433
SARS-CoV-2      FSTFKCYGVS PTKLNLCFT NVYADSFVIR GDEVQIAPG QTGVIADYNY KLPDDFTGCV 433
                ***** :*****: *****: *:***** **.****** *****:

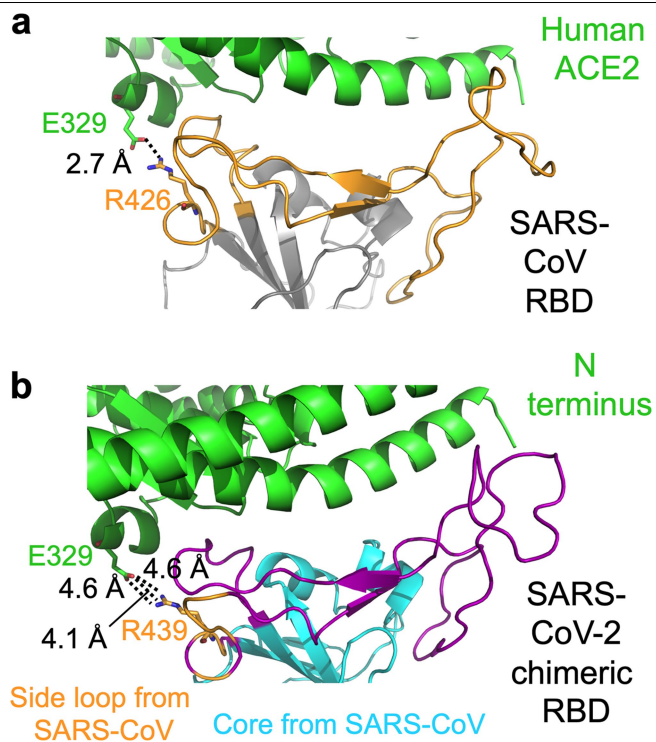
SARS-human      LAWNRNIDA TSTGNYNYKY RYLRHGKLRP FERDISNVFP SPDGKPCPT-P ALNCYWLPLND 480
SARS-civet      LAWNRNIDA TSTGNYNYKY RYLRHGKLRP FERDISNVFP SPDGKPCPT-P ALNCYWLPLKD 480
CoV-pangolin/GD IAWNRNIDS KVGGNYNLY RLFRKSNLKP FERDISTEYIY QAGSTPCNGVE GFNCYFPLQS 494
CoV-pangolin/GX IAWNRNIDS LTGDNYGYLY RLFRKSNLKP FERDISTEYIY QAGSTPCNGQV GLNCYFPLER 494
Rs3367-bat      LAWNRNIDA TSTGNYNYKY RYLRHGKLRP FERDISNVFP SPDGKPCPT-P AFNCYWLPLND 481
RaTG13-bat      IAWNRNIDA KEGGNFNYLY RLFRKANLKP FERDISTEYIY QAGSKPCNGQT GLNCYFPLYR 494
SARS-CoV-2      IAWNRNIDS KVGGNYNLY RLFRKSNLKP FERDISTEYIY QAGSTPCNGVE GFNCYFPLQS 494
                :***:..*! ..:!:*,* *:!:*,* *****:.. :!:.!*,* . :!*,*..

SARS-human      YGFYTTTGIG YQFYRVVVL FELLNAPATV CGPKL 515
SARS-civet      YGFYTTTGIG YQFYRVVVL FELLNAPATV CGPKL 515
CoV-pangolin/GD YGFHPTNGVG YQFYRVVVL FELLNAPATV CGPKQ 529
CoV-pangolin/GX YGFHPTNGVN YQFYRVVVL FELLNAPATV CGPKL 229
Rs3367-bat      YGFYITNGIG YQFYRVVVL FELLNAPATV CGPKL 516
RaTG13-bat      YGFYPTDGVG YQFYRVVVL FELLNAPATV CGPKK 529
SARS-CoV-2      YGFQPTNGVG YQFYRVVVL FELLHAPATV CGPKK 529
                ***:.*.!: :*:***** *****:*****

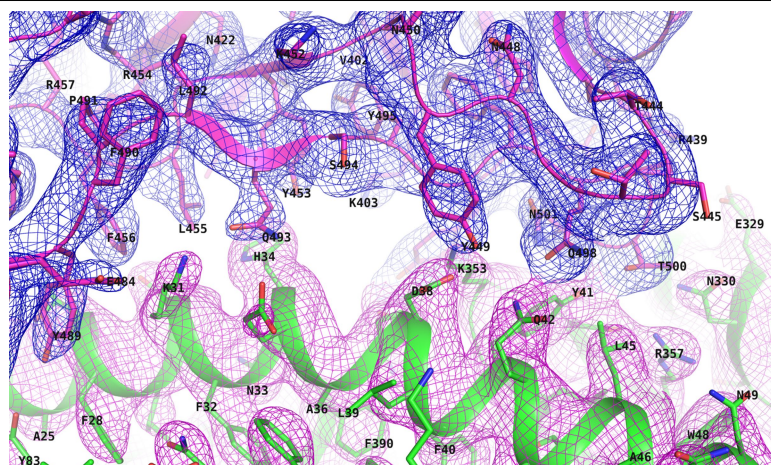
```

Extended Data Fig. 1 | Sequence alignment of the RBDs from SARS-CoV and SARS-like viruses. RBM is shown in magenta. Previously identified critical ACE2-binding residues are shown in blue. The seven RBM residues that differ between the SARS-CoV-2 wild-type RBD and SARS-CoV-2 chimeric RBD are shaded. A critical arginine on the side loop of the SARS-CoV RBM that forms a strong salt bridge with ACE2 is shown in green. Another arginine in the core structure that interacts with glycan is shown in cyan. The residues on the variable loop between two disulfide-bond-forming cysteines in the

ACE2-binding ridge are shown in red. The important motif changes in the ACE2-binding ridge are underlined. GenBank accession numbers are: QHD43416.1 for SARS-CoV-2 spike; AFR58742 for human SARS-CoV spike; AY304486.1 for civet SARS-CoV spike; MG916091.1 for bat Rs3367 spike; QHR63300.2 for bat RaTG13 spike. Two coronaviruses, CoV-pangolin/GD and CoV-pangolin/GX, were isolated from pangolins at two different locations in China, Guangdong and Guangxi; their RBD sequences were from a previous study²².



Extended Data Fig. 2 | Interface between SARS-CoV-2 or SARS-CoV RBD and ACE2. **a**, The interface between the SARS-CoV RBD and ACE2, showing a strong salt bridge between Arg426 on the side loop of the RBD and Glu329 of ACE2. The core structure is shown in grey. The RBD is shown in orange. **b**, The interface between the SARS-CoV-2 chimeric RBD and ACE2, showing a weaker, but still energetically favourable, N–O bridge between Arg439 on the side loop of the RBD and Glu329 of ACE2. The interaction between Arg439 on the side loop of the RBD and Glu329 of ACE2 is non-natural in SARS-CoV-2 (and is a result of the design of the SARS-CoV-based chimaera).



Extended Data Fig. 3 | Composite omit map of the interface between the SARS-CoV-2 RBM and ACE2. Contour level is 1σ .

a

Buried surface (Å²)	Complex 1	Complex 2
SARS-CoV-2	895.9	860.9
SARS-CoV-2 (chimeric)	924.2	883.5
SARS-CoV	849.2	829.7

c

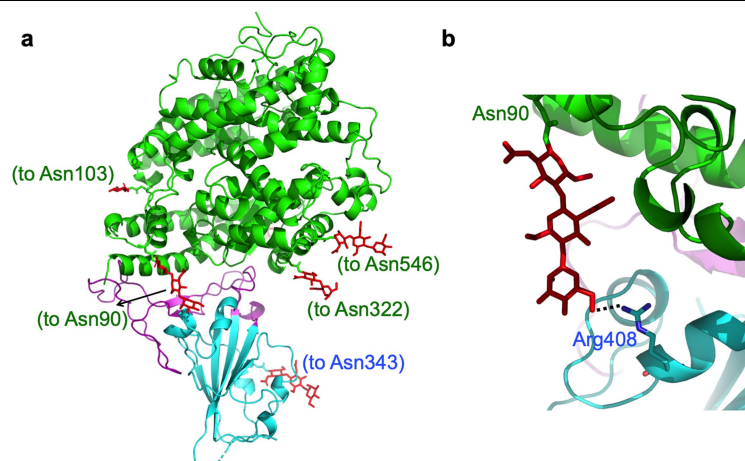
RBD	K _d (nM)	k _{off} (s ⁻¹)	k _{on} (M ⁻¹ s ⁻¹)
SARS-CoV-2 (wild type)	44.2	7.75 × 10 ⁻³	1.75 × 10 ⁵
SARS-CoV-2 (chimeric)	23.2	4.23 × 10 ⁻³	1.82 × 10 ⁵
SARS-CoV	185	3.70 × 10 ⁻²	2.01 × 10 ⁵

b

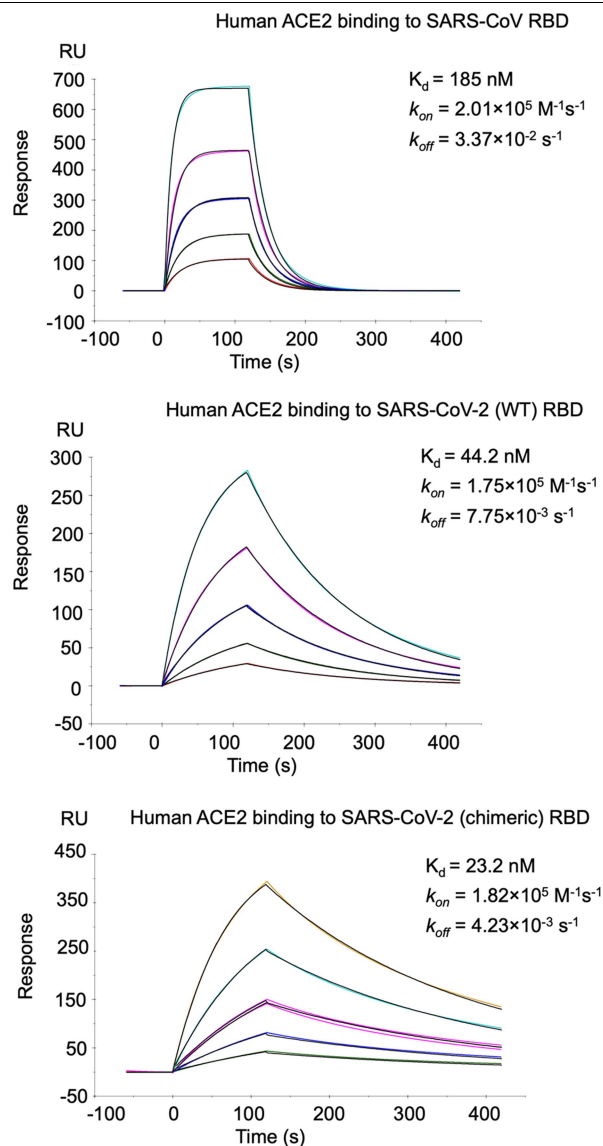
19	24	27	28	31	34	35	37	38	41	42	45	79	82	83	325	329	330	353	354	355	357	human ACE2
S	Q	T	F	K	H	E	E	D	Y	Q	L	L	M	Y	Q	E	N	K	G	D	R	
N473 Y475 Y475 Y442 Y440									Y436 Y484 Y436 Y484				L472 N473 I489 R426 T486 G482 G488 T486 T486 SARS									
Y475 N479									T486 Y484 T487				G488 Y491 T487 Y491									
A475	G476	F456	Y489	F456	Y453	Q493	Y505	Y449	Q498	Q498	Q498	F486	F486	F486	N487	R439	T500	G496	G502	T500	T500	SARS-2
N487 A475 Y489									T500 N501				N501 Y505 G502 Y505									
													chimeric									

Extended Data Fig. 4 | Comparison of ACE2 binding by the SARS-CoV RBD, SARS-CoV-2 wild-type RBD and SARS-CoV-2 chimeric RBD. a, Buried surface areas at SARS-CoV RBM–ACE2 and SARS-CoV-2 RBM–ACE2 interfaces. In the crystals for both the SARS-CoV RBD–ACE2 complex and chimeric RBD–ACE2 complex, two copies of each complex were present in one asymmetric unit. Numbers for both copies of the complexes are shown. The interaction between

Arg439 on the side loop of the RBM and Glu329 of ACE2 was excluded from the calculation of the buried surface area of SARS-CoV-2. **b,** List of contact residues from RBM and ACE2 that are directly involved in RBM–ACE2 binding. The engineered Arg439 in the chimeric RBD is shown in orange. Contact residues of the SARS-CoV RBM–ACE2 complex are taken from PDB 2AJF. **c,** Binding affinities between the RBDs and ACE2 measured using SPR.



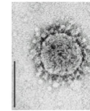
Extended Data Fig. 5 | Glycans built into the SARS-CoV-2 chimeric RBD-ACE2 structure. a, Distribution of glycans in the structure. Glycans are shown in red. The residues to which the glycans attach are indicated in parentheses. **b,** Interaction between a glycan attached to Asn90 of ACE2 and Arg408 from the RBD.



Extended Data Fig. 6 | Measurement of the binding affinities between the RBDs and ACE2 by SPR assay using Biacore. Purified recombinant RBDs were covalently immobilized on the sensor chip through their amine groups and purified recombinant ACE2 flowed over the RBDs. ACE2 was diluted to different concentrations (from 5 to 80 nM for SARS-CoV-2 RBD and chimeric RBD, and from 20 to 320 nM for SARS-CoV RBD) before being injected. The resulting data were fit to a 1:1 binding model. Each experiment was repeated independently twice with similar results. Each time, five different protein concentrations were used to calculate the K_d values.

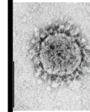
SARS-CoV-2: Where did it come from?

Bats: the natural
reservoir?



SARS-CoV-2 RBM residues	Adapted to what human ACE2 residues?
GVEG (482-485)	N-terminal helix of human ACE2
F486	M82 in human ACE2
Q493/L455	K31 and E35 in human ACE2
N501	K353 in human ACE2

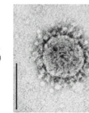
**Evolution
of RBM**



Possible interim reservoir

2019-?:
Over 2 million
infections; over 100,000
fatalities; highly
infectious

Humans



Extended Data Fig. 7 | Summary of ACE2 adaptation and evolution of SARS-CoV-2. Several structural features of the SARS-CoV-2 RBM contribute favourably to the ability of the virus to bind to human ACE2. Each of these

structural features of the SARS-CoV-2 RBM matches well with one or more structural features of human ACE2. This figure establishes the correlations among these structural features of SARS-CoV-2 RBM and ACE2.

Extended Data Table 1 | Crystallization data collection and refinement statistics

Data collection	
Space group	P12 ₁ 1
Unit cell dimensions	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	80.435, 118.034, 112.075
α , β , γ (°)	90, 93.12, 90
Resolution (Å)	59–2.68 (2.78–2.68)*
<i>R</i> _{sym} or <i>R</i> _{merge}	0.0807 (1.47)
<i>I</i> / σI	12.08 (1.12)
Completeness (%)	98.96 (98.97)
Redundancy	3.9 (4.0)
Refinement	
Resolution (Å)	59–2.68 (2.78–2.68)*
No. reflections	58219 (5774)
<i>R</i> _{work} / <i>R</i> _{free}	0.197/0.228
No. atoms	13180
Protein	12782
Ligand/ion	372
Water	26
<i>B</i> -factors	108.85
Protein	108.05
Ligand/ion	138.01
Water	82.32
R.m.s. deviations	
Bond lengths (Å)	0.003
Bond angles (°)	0.60

Data processing was carried out using HKL2000²⁵. Molecular replacement and model refinement were performed using PHENIX and CCP4^{26,27}. Model building was carried out using COOT²⁸. Structural figures were made using PYMOL (The PyMOL Molecular Graphics System, v.2.0 Schrödinger). We used 26 crystals for X-ray data collection. Each crystal resulted in one set of X-ray data. The best dataset (as judged by data statistics) was used for structure determination and refinement.

*Values in parentheses are for the highest-resolution shell.

Extended Data Table 2 | Summary of spike-ACE2 binding affinities measured by different studies

Protein coated	K_d (M)	Coating method	Detection method	References
SARS-CoV-S1-Fc tag	1.7×10^{-9}	Covalently immobilized (via amine group) to sensor chip	SPR	29
SARS-CoV-RBD-Fc tag	1.62×10^{-8}	Non-covalently immobilized (via Fc tag) to sensor chip	SPR	30
SARS-CoV-RBD-His tag	1.52×10^{-7}	Covalently immobilized (via amine group) to sensor chip	SPR	12
Human ACE2-His tag	2.09×10^{-8}	Covalently immobilized (via amine group) to sensor chip	SPR	
SARS-CoV-RBD-His tag	3.258×10^{-7}	Non-covalently immobilized (via His tag) to sensor chip	SPR	19
SARS-CoV-2-spike-His tag	1.47×10^{-8}	Non-covalently immobilized (via His tag) to sensor chip	SPR	
SARS-CoV-spike-His tag	7.7×10^{-9}	Non-covalently immobilized (via His tag) to sensor chip	Blitz	16
SARS-CoV-2-spike-His tag	2.9×10^{-9}	Non-covalently immobilized (via His tag) to sensor chip	Blitz	
SARS-CoV-S1	Similar binding affinity	Serial dilution coated on plates	ELISA	31
SARS-CoV-2-S1		Serial dilution coated on plates	ELISA	
SARS-CoV-RBD-His tag	1.85×10^{-7}	Covalently immobilized (via amine group) to sensor chip	SPR	Current study
SARS-CoV-2-RBD-His tag	4.42×10^{-8}	Covalently immobilized (via amine group) to sensor chip	SPR	

Protein-protein binding affinities are more accurately measured using SPR than using enzyme-linked immunosorbent assay (ELISA)^{12,16,19,29-31}, as ELISA often causes artefacts in protein binding³². K_d values measured using SPR depend on how the proteins are coated. Non-covalently immobilized proteins using Fc or His tags (on the opposite side to ligand-binding sites) have the advantage over covalently immobilized proteins using amine groups because the former have the ligand-binding sites fully exposed. However, non-covalently immobilized proteins risk dissociating from sensor chips, leading to under-evaluated K_d values. Covalently immobilized proteins using amine groups do not dissociate from sensor chips, but they are attached to sensor chips in many orientations; for some of these orientations, the ligand-binding sites are not approachable, leading to under-evaluated K_d values. Compared with large proteins, the ligand-binding sites on covalently immobilized small proteins are more likely to be buried, leading to under-evaluated K_d values. Compared with RBD-ACE2 binding, the spike protein-ACE2 binding is more complex: the RBD in the spike switches between standing up (to expose the RBM for ACE2 binding) and lying down (to hide the RBM) conformations^{16,19}, complicating the interpretation of measured K_d values. Therefore, K_d values measured by different SPR studies vary, depending on which protein is coated as well as the size and shape of proteins. In a previous study¹², the K_d value was higher when the RBD was coated than when the ACE2 was coated. In the present study, we could not coat ACE2 because ACE2 dissociated from sensor chips in regeneration buffer (for unknown reasons). We therefore coated the RBD, and the measured K_d value was comparable to that from the previous study¹².

Extended Data Table 3 | Critical ACE2-binding residues in SARS-CoV-2 and SARS-CoV RBMs

Viral RBD	Year	442	468-471	472	479	480	487
SARS-human	2002	Y	P-PA	L	N	D	T
SARS-civet	2002	Y	P-PA	L	K	D	S
CoV-pangolin/GD	2020	L (455)	GVEG (482-485)	F (486)	Q (493)	S (494)	N (501)
CoV-pangolin/GX	2020	L (455)	GQVG (482-485)	L (486)	E (493)	R (494)	T (501)
Rs3367-bat	2013	S (443)	P-PA (469-472)	F (473)	N (480)	D (481)	N (488)
RaTG13-bat	2020	L (455)	GQTG (482-485)	L (486)	Y (493)	R (494)	D (501)
SARS-CoV-2	2019	L (455)	GVEG (482-485)	F (486)	Q (493)	S (494)	N (501)

Residues in the SARS-CoV-2 RBM are labelled in red.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Beam line at Advanced Photon Source beamline 24-ID-E is controlled by in house developed "Console 6.2.0" suite of programs. Automated data processing is enabled by locally developed software suite called RAPD.

Data analysis

HKL2000, CCP4 7.0, PHENIX-1.17.1, PyMol 2.0, LigPlot+ program and PDBePISA web server Ver.1.48

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Coordinates and structure factors have been deposited to the Protein Data Bank with accession number 6VW1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences
- ☐ Behavioural & social sciences
- ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. For the protein expressions in insect cells, 2 liters cell culture (about 2-3x10 ⁶ cells/ml) were used each time.
Data exclusions	No data were excluded from the analyses
Replication	We have successfully repeated the crystallization condition more than 20 times. Pull-down assay and pseudovirus assay were each repeated 3 times.
Randomization	Randomization was not relevant to our study. Because there's no allocation of samples/organisms/participants involved in our study.
Blinding	Investigators were not blinded to group allocation during data collection and/or analysis. Because there's no group allocation involved in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		

Antibodies

Antibodies used	Primary antibody for C9 tag detection: rhodopsin (1D4). Its supplier: Santa Cruz Biotechnology. Its catalog number: sc-57432. Its clone name: 1D4. Its lot #: E0819. Primary antibody for HIV-1 p24 detection: HIV-1 p24 (24-4). Its supplier: Santa Cruz Biotechnology. Its catalog number: sc-69728. Its clone name: 24-4. Its lot #: F1417. Peroxidase-conjugated secondary antibody was also used for Western blotting (WB). Its supplier: Jackson ImmunoResearch. Its catalog number: 115-035-062. Its lot #: 139773
Validation	Anti-rhodopsin Antibody (1D4) is a mouse monoclonal IgG1, which is recommended for detection of rhodopsin of mouse, rat and human origins by WB, IP, IF, IHC(P) and ELISA; also reactive with additional species, including bovine. The dilution ratio is 1:1,000 for WB. Anti-HIV-1 p24 Antibody (24-4) is a mouse monoclonal IgG2b which is recommended for detection of Gag p24 of HIV-1 origin by WB, IP, IF and FCM. The dilution ration is 1:1,000 for WB. Peroxidase-conjugated secondary antibody is a goat anti-mouse IgG (H+L) which is recommended for WB with a dilution ratio of 1:10,000 - 1:20,000.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	sf9 insect cells were purchased from ATCC (ATCC® CRL-1711™). HEK293T cells were purchased from ATCC (ATCC® CRL-3216™). ESF 921 Insect Cell Culture Medium were purchased from Thermofisher Scientific (catalog #: 96-001-01). DMEM (Dulbecco's Modified Eagle Medium) were purchased from Gibco (catalog #: 11965092).
Authentication	Cell lines used were not authenticated
Mycoplasma contamination	Cell lines used were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used

Author Correction: Maximizing and stabilizing luminescence from halide perovskites with potassium passivation

<https://doi.org/10.1038/s41586-020-2272-2>

Correction to: *Nature* <https://doi.org/10.1038/nature25989>

Published online 22 March 2018



Check for updates

Mojtaba Abdi-Jalebi, Zahra Andaji-Garmaroudi, Stefania Cacovich, Camille Stavrakas, Bertrand Philippe, Johannes M. Richter, Mejd Alsari, Edward P. Booker, Eline M. Hutter, Andrew J. Pearson, Samuele Lilliu, Tom J. Savenije, Håkan Rensmo, Giorgio Divitini, Caterina Ducati, Richard H. Friend & Samuel D. Stranks

In this Letter, there are errors in several of the figures. We regret that a number of unfortunate formatting issues arose in the final submission when exporting our figures to high-resolution versions and that we did not notice these at the proof stage. These issues led to irregular spacing of tick marks, distorted symbols and lines, modified colours, and data continuing outside the axes, or becoming detached from the axes, in some line and scatter plots. Figures 1, 4 and 5 and Extended Data Figs. 1–5 and 8–11 were affected by these issues. In addition, there were some errors that propagated through from our initial submission. The colours for the data shown at the 20-, 25- and 30-minute snapshots for the $x=0.4$, $y=0.4$ composition in Fig. 2d do not correspond correctly to the colours in the common legend in Fig. 2c. Furthermore, the x -axis tick marks for Fig. 4a–c are shifted such that $t=0$ does not correspond to the initial pulsed excitation arrival time (the peak photoluminescence intensity). Extended Data Fig. 4d contains null data values beyond 2,700 ns for the $0.15 \mu\text{J cm}^{-2}$ measurement, which had been incorrectly included and scaled using the normalization procedure for the data. Finally, the plot in Extended Data Fig. 10d is missing the first three data points for the forward scan and instead contains an erroneous first data point (a legacy from a dataset analysed earlier) owing to an error in inserting the data into the plotting software. We note that all of the other data points are correct and that this error does not change the extracted device parameters. Because it is not possible to correct the original Letter online, the Supplementary Information to this Amendment contains the corrected figures.

We emphasize that these issues do not change our conclusions or interpretations in any way. We apologise that these regrettable errors were not noticed before final publication. The original Letter has not been corrected online.

Supplementary Information is available in the online version of this Amendment.

Author Correction: Tumours with PI3K activation are resistant to dietary restriction

<https://doi.org/10.1038/s41586-020-2215-y>

Correction to: *Nature* <https://doi.org/10.1038/nature07782>

Published online 11 March 2009

 Check for updates

Nada Y. Kalaany & David M. Sabatini

In Fig. 3a of this Article, the western blot at the top right (P-S473 Akt, DLD-WT, 24 h) was a duplicate of the blot on the left (1 h). This error does not affect the conclusions of the paper. The corrected panel is shown as Fig. 1 of this Amendment, together with the incorrect published panel, for transparency to readers. See Supplementary Information to this Amendment for the raw data for the corrected Fig. 3. The original Article has not been corrected.

Supplementary information is available in the online version of this Amendment.

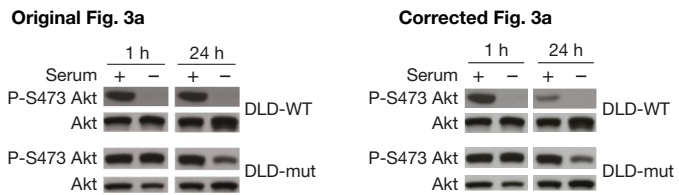


Fig. 1 | This is the corrected Fig. 3a panel of the original Article, together with the incorrect, published Fig. 3a panel.

Publisher Correction: Tunable correlated Chern insulator and ferromagnetism in a moiré superlattice

<https://doi.org/10.1038/s41586-020-2237-5>

Correction to: *Nature* <https://doi.org/10.1038/s41586-020-2049-7>

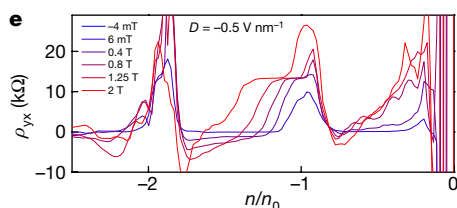
Published online 4 March 2020

 Check for updates

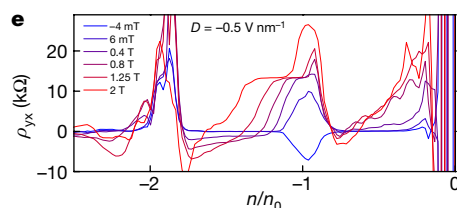
Guorui Chen, Aaron L. Sharpe, Eli J. Fox, Ya-Hui Zhang, Shaoxin Wang, Lili Jiang, Bosai Lyu, Hongyuan Li, Kenji Watanabe, Takashi Taniguchi, Zhiwen Shi, T. Senthil, David Goldhaber-Gordon, Yuanbo Zhang & Feng Wang

In Fig. 2e of this Article, the curve at -4 mT was missing, and in Fig. 2f, several of the curves were missing. Figure 1 of this Amendment shows the incorrect and corrected panels, for transparency. In Fig. 2g, the tick marks on the x axis of the inset should be labelled 0.0 T, 0.1 T and 0.2 T (not 1 T, 2 T and 3 T). In the 'Peer review information' section, the name of reviewer Fan Zhang was inadvertently misspelled 'Fan Zheng'. These errors have been corrected online.

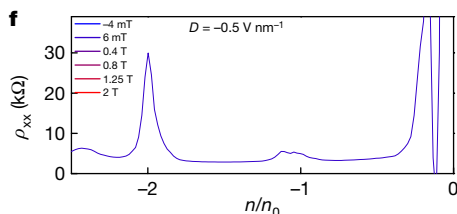
Incorrect, as-published Fig. 2e



Corrected Fig. 2e



Incorrect, as-published Fig. 2f



Corrected Fig. 2f

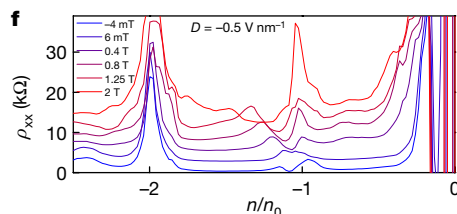


Fig. 1 | This figure shows the incorrect, as-published original Fig. 2e, f and the corrected Fig. 2e, f.

Editorial Expression of Concern: Quantized Majorana conductance

<https://doi.org/10.1038/s41586-020-2252-6>

Addendum to: *Nature* <https://doi.org/10.1038/nature26142>

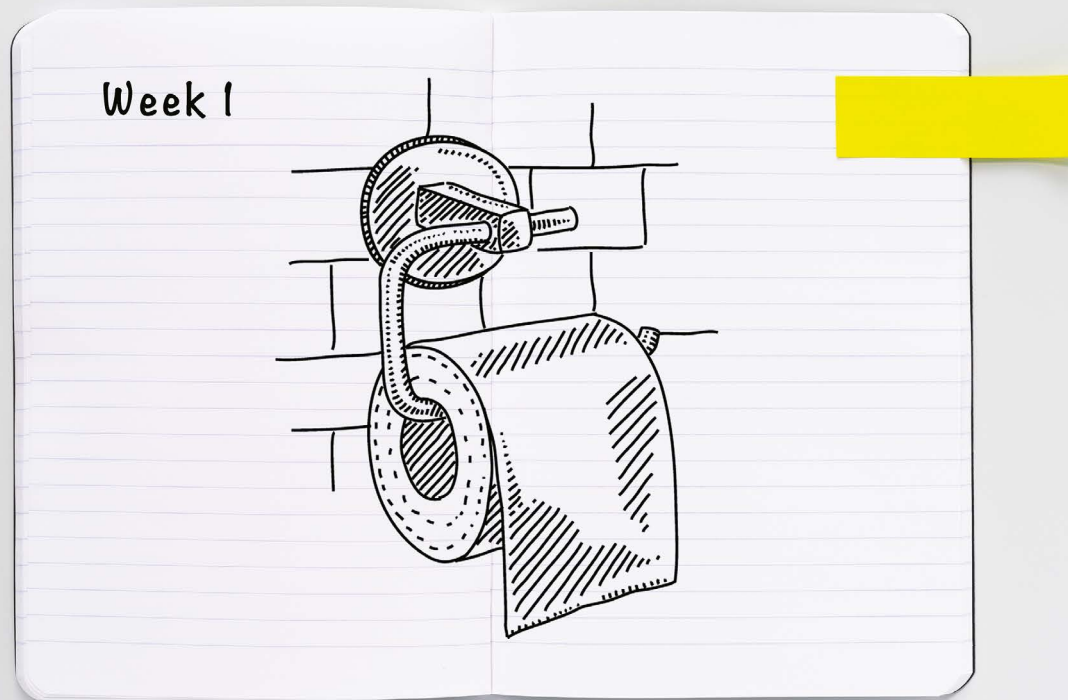
Published online 28 March 2018



Check for updates

Hao Zhang, Chun-Xiao Liu, Sasa Gazibegovic, Di Xu, John A. Logan, Guanzhong Wang, Nick van Loo, Jouri D. S. Bommer, Michiel W. A. de Moor, Diana Car, Roy L. M. Op het Veld, Petrus J. van Veldhoven, Sebastian Koelling, Marcel A. Verheijen, Mihir Pendharkar, Daniel J. Pennachio, Borzoyeh Shojaei, Joon Sue Lee, Chris J. Palmstrøm, Erik P. A. M. Bakkers, S. Das Sarma & Leo P. Kouwenhoven

The authors have alerted the editors of *Nature* to potential problems in the manner in which the raw data in this Letter have been processed, and these will have an impact on the conclusions that can be reliably drawn. *Nature* is working with the authors to resolve the matter, but in the meantime, readers are cautioned against using results from this Letter.



ILLUSTRATIONS ADAPTED FROM GETTY

CORONAVIRUS DIARIES: A NEW WORLD OF WORK

John Tregoning's weekly COVID-19 diary entries reveal the highs and lows of remote working while homeschooling two children.

Twenty years after starting my PhD and 12 years into running my own research group, focusing on respiratory infections, at Imperial College London, I would describe myself as a mid-career scientist. With the job title of reader (somewhere between an associate and full professor), I'm seen as impossibly old by my PhD students, and as a relative newbie by the more senior profs.

Of course, that was under normal circumstances.

To help navigate through the current experience, I started a diary, which *Nature* agreed to publish. I hope that my experience will be similar to that of other scientists – not those on the front line in the wards, or

developing diagnostics, vaccines and cures, but the vast majority of us, at every level of career, who are now adjusting to life away from the laboratory. Hopefully it will make someone, somewhere, smile and realize we are all going through versions of the same experience.

You can read all of my diary entries at [nature.com/careers](https://www.nature.com/careers)

WEEK ONE HELLO FROM HOME

I realize that, compared with many scientists and many more members of the public worldwide, I am in a privileged position. I'm a tenured principal investigator, so most likely will have

a job to return to, and I have a house – with a garden – which holds more than three rolls of toilet paper. But the uncertainty is unsettling as we step into the unknown with all kinds of issues: what happens to the economy and what that might mean for science; the likelihood of getting a nasty infection; uncertainty over when I will be able to buy more bread flour.

In addition to being an academic, I am a father to children aged 10 and 12, and a husband to my wife, who is also a scientist. So, for both of us, one of the biggest challenges is going to be how to fit full-time work around looking after the children. Under normal circumstances, parenting is mostly logistics. We just about balance childcare through a combination of school, clubs and childminders. This

Work/Careers

juggling act gives us enough time to fit in most of a week's work.

Given that my family is all now at home more or less 24/7, apart from the time allowed for exercise in the UK government's guidelines, things are going to have to change. In future columns, I will try to chart the changes I go through and identify what, if anything, has worked in balancing home, career and personal life. But I'll also aim to share the things that haven't. I'm already keen to write about the loss of identity I've felt with closing my lab, how I'm trying to maintain a healthy relationship with social media, the challenges of remote supervision and how to make the tallest Lego tower to win the school competition. In addition to my home family, I still have a responsibility to my work family – the students and staff members who work in my lab – and I'm still working out the best way of maintaining contact and keeping the science going.

One of the things I love about academia is the predictable patterns. The year starts in October; grant opportunities come and go. Having never really worked full-time anywhere other than academia, not counting the summer job in the refrigerated warehouse, the idea of doing anything different is terrifying. But here is an opportunity, unsought. Maybe we can all try different ways of working and being with our families. Then again, it might just be awful. Like everyone else, I'm still working it out.

WEEK TWO TO BE A SCIENTIST



For me, one of the biggest things to come to terms with, as I'm locked down in the United Kingdom, is not being able to go to my lab or my office. I realized the extent to which I was missing work when I told my children to get the ice cream from the freezer in the lab, actually meaning the garage – my subconscious speaking volumes.

Shutting the lab down came as a bit of a shock, despite the warning signs from other countries' responses to the coronavirus outbreak, and the increasingly grim news from the epidemiology modellers downstairs. I'd done some preparation the week before – mostly

making plans with my PhD students and lab technicians about where they might best see out the next few weeks (at home with family or in their London flats). But there were several unanswered questions causing me angst:

- What were my lab-facing team members going to do with their time?
- What was I going to do with my time?
- Who was going to water my office plants?

The real challenge, though, is deeper than working out what to do with my and my teams' working hours. It revolves around personal identity. So much of how I see myself is tied up with what I do as a job. I am a father, a husband, a brother. But I'm also a scientist and an academic. One of the great things about being a scientist is the close overlap between job and personal interests. But there can be times when the close relationship between science and self gets out of kilter and science takes over. There are waves of intensity, normally peaking around the time of grant deadlines, when I can think of little else.

Now, however, I'm in new territory. Not having a lab to go to will have an impact on more than just work productivity. It isn't necessarily just lab work that will be affected – I am the first to admit that I am not in the lab itself very much during the week. Like most principal investigators, I spend much of my time working on the leadership, funding and administrative tasks that spring up around wet-lab work – but the proximity to it and the interactions with my team in the lab are all part of the job. Working from home occasionally was an excellent way to get a piece of focused thinking done, but the appeal soon disappears when it is the only option.

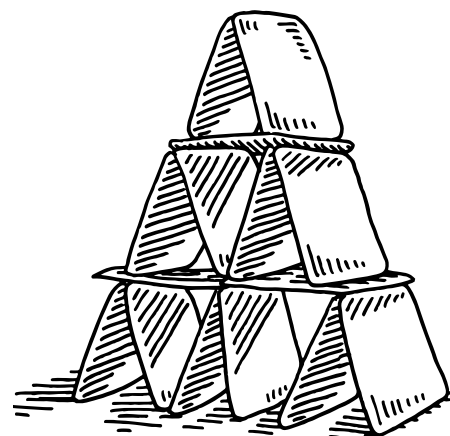
On reflection, I would, in part, link my identity as a scientist to the discovery of new things, or at least living vicariously through the work my wonderful team does. In her fantastic book *Lab Girl*, Hope Jahren describes the moment she made her first discovery and how this led her to an academic career. When I was doing my BSc, my supervisor pointed out that I was the first person ever to see what I was seeing down the microscope. It wasn't anything spectacular, but it was enough to get me hooked. These little moments of discovery are hard to achieve from my home office.

On deeper reflection, it isn't only the discoveries that drive me, but the story-telling that builds from them – stringing individual discoveries into an epic scientific tale. As my team would attest, there isn't always a solid plan at the beginning; the science builds organically from one point to the next. Each experiment leads to the next: a thread running through them from beginning to end. When your approach to planning work depends on the experiments, it is bloody hard to plan the next step without the experiments.

I need to remind myself that the current situation is temporary, if open-ended. I have had a similar challenge before: in parallel with doing my PhD and postdoc, I was an officer in the Army Reserves for ten years, and when I left, I had the feeling 'if I am not an army officer, what am I?' Turns out, it wasn't such a big deal, I was still me, even if I wasn't marching up and down the parade square. I imagine the experience was not dissimilar to retiring – which is probably why so few academics do actually retire.

And of course, I'm sure that once I'm back in the lab, the excitement will fade with the first failed polymerase chain reaction.

WEEK THREE SCHOOL'S OUT FOREVER



For the record, the three people whom I would choose to be stuck with in the same house for 4 weeks (and counting) are my wife and two children (aged 10 and 12), and if it was just a prolonged holiday with no responsibilities, it would be wonderful. Unfortunately, we all still have work to do – whether it's grants to write or a space-exploration school curriculum to study. I've enjoyed spending more time with the children since lockdown began, but being in the same space as them for four weeks has required even more juggling.

I had thought that being a working parent was tough under normal circumstances: it involves a lot of planning. My wife and I try to keep our diaries coordinated and outsource the tasks we can, by getting help from my wife's personal assistant, our cleaner, childminders and our kids' school. We are very fortunate that we normally have this support, giving us the time and space to focus on work and the more fun aspects of parenting.

You might be reading about people who seem to be effortlessly balancing the house-of-cards of home-working and homeschooling: up with the YouTube fitness coach Joe Wicks; story time with the author David Walliams; a quick trip to the virtual zoo, while baking, cleaning and grant writing. The reality for me is rather different.

Let me tell you about last week:

Monday – got up at 7 a.m., worked for 2 hours. Cooked breakfast, dodged in and out of a videoconference while answering questions from my 10-year-old about the planets.

Tuesday – carefully mapped the space-exploration curriculum into 30-minute blocks, with gaps for enrichment. Stuck to the timetable for 15 minutes of the first period.

Wednesday – bought more Wi-Fi routers, and downloaded many apps.

Thursday – began to find a pattern that worked, by creating a more flexible timetable for each child, with more breaks for electronic leisure for everyone (adults included).

Friday – had half a group meeting, failed to turn my laptop camera on throughout, which was lucky because I was also trying to teach Year-8 physics and make a papier-mâché space rocket at the same time.

Saturday – oh God, the weekend is just the same as the week, but now there aren't even school timetables to follow.

Sunday – gave up; ate biscuits; drank wine.

Monday – rinse and repeat.

This is a really challenging time to be a parent. It is challenging to be anyone. The systems that most of us have just about got in place are disrupted. We are in a time of feast or famine. Some scientists are able to go to work and are busier than ever, dropping everything to develop diagnostics, vaccines and therapeutics. And then there are the rest of us with nowhere to go, and either having nothing to do or desperately trying to squeeze work into the quiet time between the beginnings and ends of children's TV programmes. And it isn't only work: some of us are seeing way too much of our families, and others nothing at all. The current situation is not normal for anyone.

My new role as homeworker and homeschooler has certainly confirmed a few things. I love my kids, but I do not want to be their teacher. I am very grateful to the people who put all that time and effort into educating children around the world, particularly mine. I am hoping that my life as a teacher of small people will be short-lived. I do some teaching, but it is minimal and is for adults who mostly want to be there – and I have control over the subject matter. I am now a slave to the school curriculum, having to answer questions on subjects that I pushed out of my brain the minute I finished my last exams on them.

If it helps, here are some things that worked: splitting the day into two blocks of 6 hours, so one person works morning and one afternoon, is better than alternating four chunks of 3 hours; curiously, it is also better than one day on, one day off. With older children, some things can be done in parallel – I can send e-mails while they are doing work set from school. Structure helps everyone – we have a timetable for holidays as well as school days. You don't have to do all of the free online

activities: if you didn't like ballet before lockdown, there is no reason to watch it now. But we've found that taking some exercise is crucial for venting our frustrations.

Ultimately, I am very grateful to my two classroom assistants Mr I. Pad and Miss X. Box, who have delivered sterling and consistent service throughout.

PS, For those of you who were worried, I have found a source of bread flour.

WEEK FOUR CREATURE COMFORTS



Like most people's, my working life at the moment is far from normal. It's not so much the working from home – it's the never leaving home. I miss the familiar surroundings of work, especially my plants. I have lived in my current house for 10 years; I have worked at Imperial College London for more than 20 years, 15 of which have been on the St Mary's campus in Paddington in Central London. And since I moved out of London to live, I've almost certainly spent more time at St Mary's than at my house (minus sleeping). Working from home and the new routines it requires have had a massive impact on my ability to concentrate.

As I have written before, a large chunk of science is creativity. This needs time and space. In the good old days, working from home meant retreating from the endless stream of meetings and interruptions, and having some space to think about work. Now, working from home is very different. When I am not making papier mâché rockets and supervising homework club for my children, there is still plenty to do. This is made much trickier when work has to contend with the sounds of something more fun than work or the smell of chickpea and chorizo soup rising from downstairs, the robin that has made our garden its home, or basically any excuse to leave my desk.

My natural response is to get distracted. Modern life exacerbates this: the ding of the phone, the notification from Twitter, the e-mail

envelope. My normal strategy is to get all of that out of the way and then focus on the work in hand – usually by allowing myself a timed 'block' of distraction before working. This varies: before writing this article, I tried to make a sourdough loaf (which turned out denser than a neutron star), watched the music video for 'Acquiesce' by Oasis and checked my son's progress in the video game *FIFA 19* (who needs real sport when you have the emotional journey of a 12-year-old trying to win a key game?). As you can see, sometimes it takes more than the five minutes I'd usually allow. However, the work blocks are now much shorter because my wife and I are rotating between work and childcare – trying to fit 50 hours of work into 20 hours – and so starting each block with 5 minutes of faffing is eating into work time. Yes, the simple solution is not to get distracted, but that is easier said than done.

It's not as if distraction was impossible when I had access to an office – there was always a member of the team to chat to, and Oasis music videos were still within reach. But there, I had worked out ways to stay focused. I could reward myself for an hour of work by getting a treat, such as a cup of tea. But this, too, is affected by being at home all the time. Previously, I could leave the house pretty messy because I knew I would not have to see it for the rest of the day, and therefore not worry too much about it. Now, to have a cup of tea in one of my too-frequent breaks, I end up emptying the dishwasher, which is always full.

I am, for what it's worth, beginning to develop some approaches that help. I've accepted that there is going to be some faffing before I get going: using my limited stock of willpower on breaking long-established habits is wasted energy. I try to cut access to my phone, ideally by leaving it in another room, but at least out of sight and reach. I try to keep my weekday routine approximately the same as before, specifically having a shower at as close to normal time as usual. I am an inveterate list maker: I have daily, weekly, monthly to-do lists. All are dauntingly long at the moment, but they still help to provide some structure and a clearer sense of what I can do with the time available. This means saying no to more things – I'm sorry to report that my contribution to peer-reviewing has been minimal of late.

In the end, it comes down to accepting that I cannot get as much done as I would if I wasn't stuck at home. But there is a positive trade-off: I am getting to spend a great deal more time with my children, which has been a real gift. One of my favourite lockdown stories is of a dog that strained its tail by wagging it too much because its owners were home all the time. Let's just say I'm glad my kids don't have tails.

John Tregoning is a reader in respiratory infections in the Department of Infectious Disease, Imperial College London.



IVAN BUSIC/SHUTTERSTOCK

Although single-use plastics are essential for certain experiments, some scientists are striving to reduce such waste.

DIY APPROACHES TO SUSTAINABLE SCIENCE

Research labs are huge producers of plastic waste, but scientists are becoming increasingly aware of their environmental footprint. **By Jyoti Madhusoodanan**

As a postdoctoral researcher, Cristina Azevedo went through single-use plastic tubes by the hundreds. The University College London biochemist was culturing yeast in Falcon tubes, and the thought of all that plastic waste was like an itch she couldn't scratch – especially when she recalled her PhD research, in which she grew bacteria in reusable glass flasks. “My own work was bothering me, and all around I could see the amount of plastic just being thrown out because of the need for sterility,” she says.

She's not alone. Scientists are increasingly aware of the disproportionate environmental footprint of their research. Academic research facilities consume three to six times as much energy as commercial buildings, much of that due to refrigeration and ventilation systems. These facilities are also outsized producers of plastic waste – an issue that has become

particularly acute since 2017, when China stopped accepting several types of plastic for recycling from the United States and Europe, causing more recyclable waste to be piled into local landfills.

At the institutional level, many facilities are stepping up, implementing better waste-management practices and seeking out greener energy sources. University College London, for example, is striving to be rid of single-use plastics by 2024 and to be carbon-neutral by 2030. Not all of these efforts will translate easily to individual laboratories, where ultracold freezers and single-use plastic pipette tips remain necessary for certain sensitive experiments. But when it comes to most standard bench science, little changes can go a long way.

Leeba Ann Chacko, junior research fellow at the Indian Institute of Science in Bengaluru,

is another yeast researcher who reduced her plastic consumption by switching to glass Petri dishes to grow her microbes. Instead of a large bag of waste each week, she now generates only a few hundred grams. “Starting out, I was worried about contamination and the expense,” Chacko says. “But it was a quick transition and left me wondering why we hadn't done this earlier.”

Beyond reducing researchers' environmental impact, such efforts can help to make their funding go further. And there are more intangible benefits, including greater reproducibility and career perks. Azevedo's sustainability work helped her CV to stand out when she applied for her current position as director of the biopesticides department at the António Xavier Chemical and Biological Technology Institute in Lisbon, for instance.

Plus, sustainability is simply responsible

science, these researchers say. “We’re being funded by public money,” Azevedo says. “We have a social responsibility to think about the environment and the planet’s future when using that money.”

Starting small

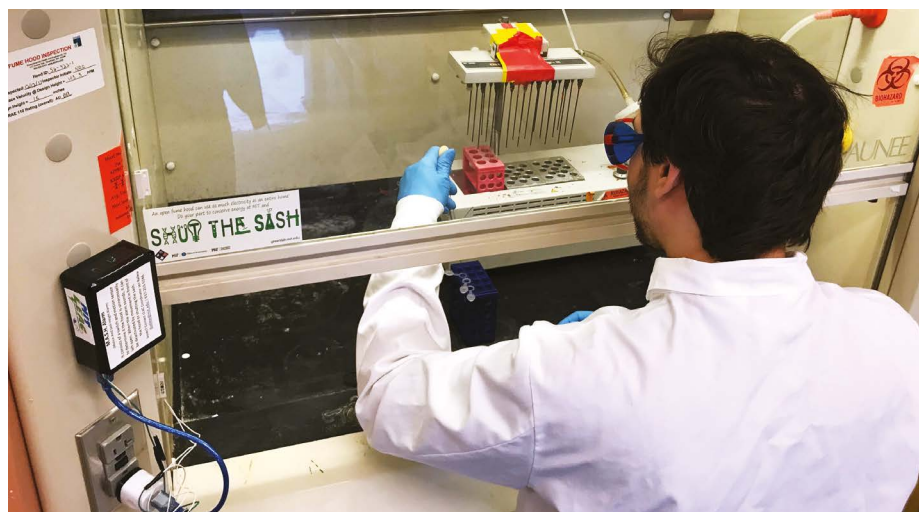
Mechanical engineer Dan Preston of Rice University in Houston, Texas, was drawn into lab energy savings during his undergraduate years, thanks to a part-time job with the Alabama Industrial Assessment Center in Tuscaloosa, where he worked with researchers who offered energy-saving recommendations to local factories.

As a postdoctoral researcher at the Massachusetts Institute of Technology (MIT) in Cambridge, he got the chance to put that knowledge into action. Preston’s research relied heavily on handling chemicals in a fume cupboard. He noticed that the sash window on the cupboard was frequently left open, which wasted energy by leaving the ventilation running, and wondered if he could engineer a simple solution. He and his colleagues entered a competition sponsored by MIT Green Labs, a programme in the environmental health and safety office that helps departments and labs across campus to operate in sustainable ways. They won around US\$5,000 to develop a simple, unobtrusive sensor, which they named the motion and sash height (MASH) alarm, that alerts users if a sash is left open. He and his team also set up the Lab Energy Assessment Center (LEAC) to evaluate labs and offer energy-saving recommendations.

One MIT lab, for instance, was spending about \$30,000 on electricity and releasing 163 tonnes of carbon dioxide per year, the LEAC team estimated. By increasing freezer temperatures from -80 to -70 °C, replacing overhead lights with LED bulbs and turning off one fume cupboard, the lab could reduce energy usage by 8% and save around 13 tonnes of CO₂ and \$2,500 annually. “Many were surprised by the impact of these small changes,” Preston says.

University funding can provide another avenue to support sustainability. David Waterman and Brenda Lemos were molecular and cell biology graduate students at Brandeis University in Waltham, Massachusetts, in 2017, when China announced it would no longer accept plastic waste from the United States and Europe. The two put together a business proposal for a recycling firm and defended it in front of a panel of venture capitalists as part of a programme called SPROUT, a university award to support student entrepreneurs. The resulting financial support and mentorship allowed the team to launch GreenLabs Recycling, a company that recycles the plastic boxes used to store single-use pipette tips.

Defending a proposal to launch a business felt like the opposite of a thesis defence,



An energy-saving alarm (shown left) signals when a fume cupboard is left open.

“where you’re the world’s expert on the research material”, Waterman recalls. “I found it much more stressful, but you do leverage the soft skills of a PhD, such as how to convincingly tell a story.”

GreenLabs Recycling now has 16 customers in the Boston area, and receives roughly 1,400 kilograms of plastic pipette-tip boxes per week. Researchers often ask how they can use the company’s services, but Lemos explains that researchers and firms first need to interact with their facilities managers to address the issue. “Although we as scientists are aware of the lack of recycling in science, facilities managers are less familiar with the problem,” Lemos says. “So there’s an initial disconnect when you’re trying to get people to pay for recycling.”

Preston and his colleagues, supported by

“Sustainability is simply responsible science.”

MIT’s environmental health and safety office, expanded LEAC to increase the number of labs it evaluated. The team also made MASH assembly instructions open-access, so anyone can build their own alarm system for about \$50. Working with undergraduate students is a big part of LEAC’s efforts, Preston says. “We think of it as saving energy, but also as inspiring and educating the next generation of energy-conscious researchers.”

Every step counts

Even researchers without the support or funding needed for such large-scale efforts can make a difference to their own lab’s energy footprint, according to Kathryn Ramirez-Aguilar, programme manager for the green labs scheme at the University of Colorado Boulder. One simple way is to share equipment.

At her university, departments and groups of researchers began that process informally,

with each lab sharing an under-utilized instrument. Some have formalized these arrangements by hiring managers to maintain shared instruments, and creating memoranda of understanding for users. The arrangements conserve research dollars and help users to share expertise on making the most of various instruments, which are housed in individual labs or a common area. They also conserve lab space, which tends to be the most expensive, energy-intensive space on campuses. “Can you imagine the impact if we didn’t have to build a whole other lab building, simply because we were using our space more efficiently?” Ramirez-Aguilar says.

That impact could extend to scientific reproducibility and efficiency by reducing the need to repeat failed experiments, argues Martin Farley, University College London’s sustainable-labs adviser. “Research requires an immense investment of energy and materials, and the data generated represents an investment of those materials,” he says. “Any way that we can promote better use of the data, or techniques that reduce error and repetition, are going to be more sustainable.”

Plus, shared instruments are likely to be better maintained than equipment in an individual lab, Farley adds. Researchers with these facilities “get better support around experimental design, and they understand better what the equipment can actually do”, he says.

The bottom line is: when it comes to sustainability, individual researchers can make a difference. And every little bit helps. “Whether you choose to champion sustainability in your own lab or whether you want to work with many different labs through something like LEAC, it takes both to implement real change,” Preston says. “You can have a huge impact either way.”

Jyoti Madhusoodanan is a science writer based in Portland, Oregon.



**The week's
best science,
from the world's
leading science
journal.**

NATURE.COM/NATURE/PODCAST

nature

A80540

Work / Technology & tools

SOUND BYTES: SIGHTLESS CODING

For visually impaired researchers, learning to program can be challenging. A tool called CuriO offers a multisensory route. **By Constance Clare**

"Technology is now essential to the lives of blind people, yet many computer programs and devices aren't universally accessible," says Jo Fullerton, Technology for Life Coordinator at the Royal National Institute of Blind People in Edinburgh, UK.

While pursuing a master's degree in integrated product design at Delft University of Technology in the Netherlands, Krishna Rajagopal developed a tool that he thinks can help. CuriO combines auditory and tactile stimuli to make coding accessible to people with visual impairments.

"CuriO brings an exciting new perspective to programming. For the first time, blind developers have the choice of a multimodal device that could make programming quicker, easier and more efficient," says Parham Doustdar, a software developer at the trip-planning website, Booking.com, who is completely blind.

Nature spoke to Rajagopal, now based in Chennai, India, to learn what makes CuriO tick.

Why did you develop CuriO?

Coding has become an essential skill of the modern world, and many employers need skilled programmers for a wide variety of roles. Although many people with limited vision aspire to code just as their sighted counterparts do, they don't have equal opportunities.

That's especially true in India, where I'm from, and other developing countries. I aimed to design an affordable and accessible tool that could close the gap, opening up lucrative careers to the visually impaired community.

How does CuriO work?

CuriO allows blind and partially sighted users to navigate and skim programming code. When code is written using an external keyboard, or is loaded into CuriO's text editor, on-screen visual elements, such as code structure and hierarchy, are translated into a pattern of moving buttons organized in rows on the device. Each row represents a line of code, and the position of the button indicates indentation. As the user navigates the code using a joystick, button positions update in real time. By pressing certain buttons, users can hear specific characters, words, lines or pages read aloud. This allows users to correctly place the cursor for debugging.

CuriO is compatible with Python, Java, Javascript and C++ code, and runs on Linux.

How is CuriO different from other products?

What sets CuriO apart is its multisensory approach to programming. Products available to visually impaired programmers are typically limited to screen readers and refreshable Braille displays. Unfortunately, screen readers are designed to read natural language rather than source code; and Braille displays are expensive and not widely accessible owing to a decline in Braille literacy. I designed an interface that combines touch and sound to make coding easier.

What's next?

In January, I brought the prototype to the I-Stem Confluence, an event in India that showcases accessible products designed for scientists, technologists, engineers and mathematicians. At this year's conference, which welcomed around 120 delegates to Bangalore, four visually impaired developers took CuriO for a test drive. I'm now looking to collaborate with industry experts to bring it to market.

Constance Clare is a science writer based in Nottingham, UK.



CuriO uses tactile and audible signals.



Where I work Jessika Trancik

I've been an energy-systems researcher at the Massachusetts Institute of Technology in Cambridge for ten years. My research helped to guide US government strategy ahead of the Paris climate-change negotiations, and has been presented to the International Energy Agency in Paris, which helps nations to shape their energy policies.

My team and I model how electric batteries might affect fossil-fuel use, and how they are becoming more affordable and efficient. This information helps governments and policymakers to assess the effects of tax rebates on electric-vehicle purchases, for example. It also helps those considering investing in carmakers, as well as engineers who are developing road networks.

In this picture, taken in late January in Cambridge, I'm holding a tablet loaded with Carboncounter, an app that tells users about vehicles' greenhouse-gas emissions and running costs. I developed it with two of my students in 2016 to help users choose vehicles with low environmental impacts.

Here, I'm matching passing cars with data in the app about emissions and costs for those models. Electric vehicles, for example, score highly on my app because of their lower emissions, but larger, boxy vehicles are less aerodynamic and thus score much worse.

Since the coronavirus pandemic, my team and I have been working from home, meeting regularly online and continuing our work.

Carboncounter is easy to use and customize, so it's popular with climate policymakers and teachers. Quebec and Ontario legislators and university educators have also used it. Legislators can use the data to create incentives for manufacturers to regulate vehicle emissions; to offer consumers rebates; and to fund research into vehicles that are more energy efficient.

I hope that some of these approaches will make high-efficiency electric vehicles more affordable for everyone.

Jessika Trancik is an energy-systems researcher at the Massachusetts Institute of Technology in Cambridge. **Interview by Sarah Boon.**

Photographed for *Nature* by
Kayana Szymczak.

nature

outlook

COPD



Revealing the true extent of
a deadly lung disease

Produced with support from:



Boehringer
Ingelheim



For more on chronic obstructive pulmonary disease visit www.nature.com/collections/COPD-outlook

Editorial

Herb Brody, Richard Hodson, Jenny Rooke

Art & Design

Mohamed Ashour, Kate Duncan

Production

Nick Bruni, Karl Smart, Ian Pope, Kay Lewis

Sponsorship

Stephen Brown, Natasha Boyd, Claudia Danci

Marketing

Nicole Jackson

Project Manager

Rebecca Jones

Creative Director

Wojtek Urbanek

Publisher

Richard Hughes

VP, Editorial

Stephen Pincok

Managing Editor

David Payne

Magazine Editor

Helen Pearson

Editor-in-Chief

Magdalena Skipper

Respiratory health has never been at the forefront of so many people's minds than it is now. The COVID-19 pandemic has brought about rapid and far-reaching changes to daily life that were inconceivable just a few months ago. It is all hands on deck in the fight against this coronavirus.

But this is not the case for other threats to our lungs. Chronic obstructive pulmonary disease (COPD) is the third leading cause of death worldwide – only coronary heart disease and stroke claim more lives each year. And yet COPD, which causes the small airways to narrow and lung tissue to break down, has long been overlooked. It is substantially underdiagnosed, even in places where prevalence is high (see page S20); has no cure; and the therapies that are available to manage symptoms are commonly repurposed treatments for asthma.

The global response to the COVID-19 crisis is unprecedented and it is unlikely that chronic threats will ever attract the same attention, no matter how large the problem. But scientists are drawing attention to the magnitude of the burden on global health that COPD represents.

They are working to improve their understanding of the disease, including the role of senescent cells – sometimes referred to as zombie cells owing to the 'undead' state in which they persist (S7). They are also investigating small packages of molecules known as exosomes. These extracellular vesicles are found in greater quantities in people with COPD, and might provide a route to new therapies (S10).

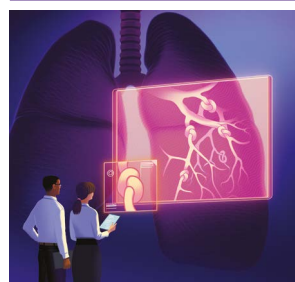
Many questions about COPD remain, such as how to protect people with the disease during wildfires (S18). But in some cases, our broadening understanding of COPD is already beginning to influence clinical practice. Some researchers are calling for the criteria commonly used to diagnose the condition to be rewritten (S4). Meanwhile, the widespread prescription of corticosteroids for COPD is being challenged (S12). And the devices used to deliver these and other inhaled therapies are changing, both to improve their effectiveness and to reduce their environmental footprint (S14).

We are pleased to acknowledge the financial support of Boehringer Ingelheim in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

Richard Hodson
Supplements editor

Contents

- S2 CLINICAL TRIALS**
Research round-up
Updates from the latest COPD studies
- S4 DIAGNOSIS**
Redefining a disease
Is it time to expand the diagnostic criteria for COPD?
- S7 CELL BIOLOGY**
Zombies in the lungs
The role of senescent cells
- S10 EXOSOMES**
Care packages
Could exosome vesicles offer possible therapy?
- S12 THERAPY**
The steroid debate
Over-prescription for COPD
- S14 DRUG DELIVERY**
The inhaler makeover
Environmental concerns are driving a change
- S18 WILDFIRES**
Fireproofing the lungs
Advice on how to keep safe during wildfires is lacking
- S20 Q&A**
Turning the tide
María Victorina López Varela explains how she and others revealed the true scale of COPD in Latin America.



On the cover

Researchers study the damage to lungs caused by COPD.
Credit: Sam Chivers

About Nature Outlooks

Nature Outlooks are supplements to *Nature* supported by external funding. They aim to stimulate interest and debate around a subject of particularly strong current interest to the scientific community, in a form that is also accessible to policymakers and the broader public. *Nature* has sole responsibility for all editorial content – sponsoring organizations are consulted on the topic of the supplement, but have no influence on reporting thereafter (see go.nature.com/2nqaz1d). All *Nature Outlook* supplements are

available free online at go.nature.com/outlook

How to cite our supplements

Articles should be cited as part of a supplement to *Nature*. For example: *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2020).

Contact us

feedback@nature.com
For information about supporting a future *Nature Outlook* supplement, visit go.nature.com/partner

Copyright © 2020 Springer Nature Ltd. All rights reserved.

Research round-up

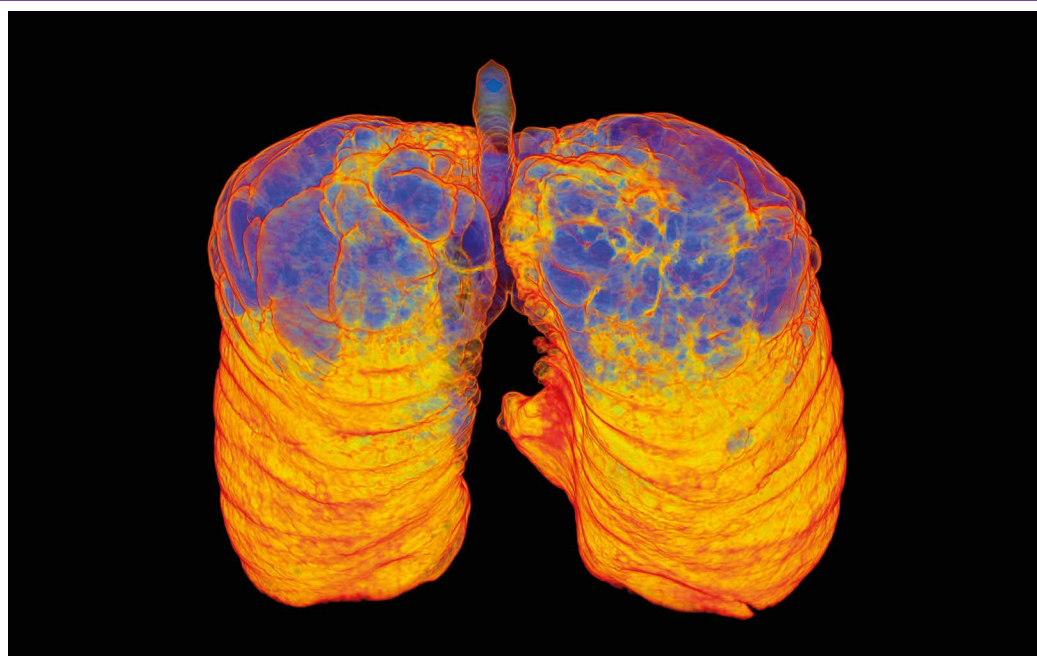
Highlights from COPD trials. By Simon Makin

Genetic risk factors uncovered

Two papers from large international collaborations have revealed genetic contributors to lung function and chronic obstructive pulmonary disease (COPD) that illuminate biological pathways and potential drug targets. They are some of the largest genetic studies of lung function and COPD so far.

The first, led by Nick Shrine at the University of Leicester, UK, assessed genetic markers linked with lung function in a genome-wide association study of about 400,000 individuals of European ancestry. The researchers identified 279 genetic regions associated with lung function, 139 of which were new. The team showed that a combination of these variants could be used to predict COPD in both smokers and non-smokers. It also found that the risk for smokers could vary by as much as fivefold.

The findings highlight the importance of several biological pathways in COPD, including the development of cilia. These tiny hair-like structures help to clear mucus from the airways, and their function is known to be impaired in COPD. The research suggests that cilia dysfunction is not merely a consequence of COPD, but a driver of the disease. The team also examined associations between specific genetic variants and a wide range of disease traits, and found evidence that COPD shares some genetic determinants with autoimmune diseases.



K. H. FUNG/SPL

Computed tomography image of lungs showing signs of emphysema (blue).

The second, which looked specifically at the genomes of people with COPD, was led by Phuwanat Sakornsakolpat at Brigham and Women's Hospital in Boston, Massachusetts. It compared data from about 36,000 people with COPD with data from some 222,000 healthy people. The researchers found 82 regions associated with COPD, some of which had not previously been linked with the disease. Of these, 22 regions have never been associated with lung function. The identification of these additional genetic loci in people with COPD suggests that there might be factors that affect the development of the disease, but that do not measurably affect lung function.

These studies underline the genetic diversity of COPD and reveal more of the genetic factors that make some people more susceptible to the disease.

Nature Genet. **51**, 481–493 (2019);
Nature Genet. **51**, 494–505 (2019)

Imaging biomarker validated in humans

A non-invasive imaging technique that can characterize different types of COPD pathology has been shown to be accurate when compared with results of tissue-sample analysis. A team has shown that the technique, based on computed tomography (CT) scanning, can differentiate disease affecting the small airways from emphysema (damage to air sacs called alveoli). Evidence suggests that damage to small airways precedes emphysema, and might accumulate for years before it is detected through a lung-function test. The new technique could identify people with early stage disease, when treatment might be the most effective.

CT scans can already detect enlarged alveoli that cause emphysema by measuring lung-tissue density when the lungs are full. Small-airway disease also alters tissue density by trapping

gas, but conventional CT does not have sufficient resolution to visualize these airways, which are about 2 millimetres in diameter. The new method, called parametric response mapping (PRM), compares CT scans conducted after inspiration and expiration to calculate changes in lung-tissue density that indicate small-airway disease. The team says PRM allows it to differentiate between density changes associated with emphysema and those that signal damage to the small airways.

Meilan Han at the University of Michigan, Ann Arbor, led the collaboration of radiologists and surgeons to test the accuracy of the technique on tissue removed from the lungs. The team conducted PRM in 11 people with severe COPD before lung transplant surgery. The researchers then used micro-CT – an ultra-high-resolution technology that cannot be used on people owing to high energy X-rays – to analyse lung tissue

samples. Regions designated as small-airway disease by PRM corresponded to characteristics of that condition – the loss, narrowing, thickening and obstruction of the bronchioles.

Future studies will need to show that PRM is accurate in milder disease. But it could be used to identify smokers at risk of developing COPD, or track the impact of new therapeutics on disease progression.

Am. J. Respir. Crit. Care Med. **200**, 575–581 (2019)

Phase III trial for nerve therapy begins

A therapy that disrupts nerves in the lungs has entered clinical trials to test its efficacy in preventing flare-ups of COPD.

Many people with COPD have periods of symptom exacerbation, which can lead to hospitalization. Targeted lung denervation (TLD) aims to reduce flare-ups by delivering a radiofrequency electrical current that disrupts nerves in the lung. The technique is based on research suggesting that the neurotransmitter acetylcholine causes airway constriction, excess mucus and inflammation in COPD. TLD is intended to produce a longer-lasting effect than anticholinergic inhalers, which deliver a chemical to block the action of acetylcholine, by disrupting the production of acetylcholine by nerves.

In a double-blind, placebo-controlled phase II trial led by Dirk-Jan Slebos of the University of Groningen, the Netherlands, the technique was shown to be safe. Half of the 82 participants with COPD received TLD therapy alongside their existing treatments. Between 3 and 6.5 months after treatment, 32% of those treated had respiratory adverse events, compared with 71% of people in the placebo group. Over a year, the risk of hospitalization was also much lower in the treatment group.

Outcomes, such as patient-reported symptoms, were not significantly different – but a phase III trial, aiming to enrol 520 people, is now under way and might prove more sensitive.

Am. J. Respir. Crit. Care Med. **200**, 1477–1486 (2019)

Blood-cell secretions spark symptoms

Small packages of molecules known as exosomes have been found to cause the damage to lung tissue seen in COPD. A study led by Edwin Blalock at the University of Alabama at Birmingham showed that exosomes released by white blood cells called neutrophils generate COPD pathology in the lungs of healthy mice.

Exosomes are secreted by all cells, and are involved in cell signalling or transporting cargo such as enzymes. Neutrophils become activated when they detect an infection. When this happens, they shed exosomes covered in the protein-degrading enzyme neutrophil elastase. In the lungs, this breaks down collagen and elastin in the extracellular matrix – the scaffold that supports the grape-like structures known as alveoli.

Ordinarily, the lungs are protected against this damage by $\alpha 1$ -antitrypsin, which inhibits neutrophil elastase. However, the team found that the enzyme escapes this inhibition when bound to an exosome. The exosomes also carry another surface protein called Mac-1 that binds to collagen fibrils, increasing the exosome's capacity to damage tissue. As a result, exosomes from activated neutrophils were found to be 10,000 times more damaging than neutrophil elastase alone.

The team showed that when activated neutrophil exosomes were collected from the lung fluid of people with COPD and transferred to mice, they enlarged alveoli and increased

airway resistance. Dislodging neutrophil elastase from exosomes, or inhibiting the enzyme or Mac-1, could be potential treatments.

Cell **176**, P113–P126 (2019)

Bioengineered lung transplants advance

Lung transplantation is a last-resort therapy for end-stage COPD. But the supply of donor organs is limited, and transplants can be rejected. Bioengineered lungs might offer a solution. A team led by Joaquin Cortiella at the University of Texas Medical Branch at Galveston has reported the most advanced attempt so far to transplant bioengineered lungs into pigs.

Previously, transplanting bioengineered lungs into rats resulted in haemorrhage, coagulation and swelling, so the team's main aim was to produce a working structure of blood vessels that could support long-term survival of transplanted tissue. To create the lungs, the team stripped pig lungs of cells, leaving the extracellular matrix, before repopulating these 'scaffolds' with cells derived from recipients' lungs. It then added plasma and proteins to develop the lungs in culture over 30 days.

Once transplanted into pigs, the lungs developed vascular tissue and blood circulation, as well as alveolar tissue. There were no issues with rejection, and the pigs survived for up to two months. The organs had similar gene expression and immune function to the pigs' own lungs, and native microorganism communities established themselves in the organs.

The lungs did not contribute to gas exchange, because the bioengineered vasculature was not linked to host arteries. Nevertheless, this is the longest an animal has survived in a study of this kind. Extracellular lung components are highly conserved between pigs and

humans, so it is possible that pig scaffolds could be used in human transplants.

Sci. Transl. Med. **10**, eaao3926 (2018)

Corticosteroids disrupt microbiome

Inhaled corticosteroids might allow bacterial infections to flourish by altering the community of microorganisms in the lung. A team led by Sebastian Johnston at the National Heart and Lung Institute, Imperial College London, showed that steroids disrupt lung microbiota by inhibiting antibacterial molecules, which might explain previous studies that indicate the drugs increase the risk of pneumonia in people with COPD. The team analysed sputum samples from people with COPD to show that inhaled steroids are associated with disrupted lung microbiota and proliferation of *Streptococcus* bacteria. When mice were given steroids, there was a similar increase in bacteria.

In human cells and mouse models, the team showed that steroids impair the clearance of *Streptococcus pneumoniae*, the most common bacterial cause of pneumonia, by suppressing the antimicrobial peptide cathelicidin. The steroids seem to increase expression of a protease that degrades cathelicidin, called cathepsin D. The team also showed that cathelicidin reversed the increased bacterial load seen in steroid-treated mice, as did inhibiting cathepsin D. These strategies could be effective in preventing or treating COPD complications.

Science Trans. Med. **11**, eaav3879 (2019)



For the latest research published by Nature visit: <http://go.nature.com/2xyawlX>

Redefining a disease

A proposal to expand the diagnostic criteria for chronic obstructive pulmonary disease puts overlooked groups of patients in the spotlight. **By Amanda Keener**



JODI JACOBSON/GETTY

Spirometers can be used to measure lung function, but results often indicate that people with the symptoms of COPD are healthy.

Pulmonologist James Crapo might be semi-retired, but that hasn't stopped him from trying to revolutionize the field of chronic obstructive pulmonary disease (COPD). At 76, Crapo remains co-director of a massive observational study of smokers across the United States called COPDGene, which he and his colleagues started 12 years ago at National Jewish Health in Denver, Colorado. Since 2008, COPDGene researchers have worked to define the spectrum of disease courses that lead to COPD by tracking the health and genetics of more than 10,000 current and former smokers. The researchers' main goal is to understand why only some people develop the disease. But along the way, the data have led them to conclude that the current definition of the disease is much too narrow. As far as Crapo is concerned, it needs to be completely rewritten.

Since the late 1990s, COPD has been diagnosed according to a set of criteria developed by the Global Initiative for Chronic Obstructive Lung Disease (GOLD). Clinicians base their diagnoses on a person's symptoms – a persistent cough, heavy mucus production and shortness of breath – as well as their exposure to risk factors such as smoking and the results of a test of lung function, called spirometry, that measures how much air a person can force out. The lung-function score must be below a certain threshold for a person to be diagnosed with COPD.

The difficulty is that there are huge numbers of people who have the symptoms of COPD, and clear signs of airway inflammation and lung damage on computed tomography (CT) imaging, but whose lung-function tests indicate that they are healthy. Within the COPDGene cohort, nearly 40% of the people

who didn't meet the definition of COPD when they joined the study had late-stage disease five years later¹.

"Many smokers are symptomatic despite a normal lung function – they should not be considered healthy," says Frits Franssen, a respiratory physician and researcher at the Maastricht University Medical Center in the Netherlands. "We all know that there are patients that have rather severe emphysema but normal spirometry, and it's a challenge to classify these patients." Without a formal diagnosis of COPD, these people are left out of clinical trials. Clinicians don't have the evidence they need to tell such patients what to expect and to choose the best treatments. Physicians usually treat the symptoms, often with the same drugs used for COPD, but without knowing what biological process they are targeting or whether the drugs will have long-term benefits.

Crapo thinks that the best way to ensure these patients are diagnosed and can take part in clinical trials is to introduce new subtypes of COPD. That requires new diagnostic criteria. In November 2019, he and around 100 other researchers proposed a revised system for COPD diagnosis that takes into account lung inflammation and tissue damage captured with CT imaging, and uses a broader definition of abnormal lung function, in addition to existing criteria of a history of smoking and displaying symptoms of the disease². The expanded criteria would increase the number of people in the United States diagnosed with COPD by 5–10 million, Crapo says.

Without evidence on how best to treat these patients, it is unlikely that GOLD will adopt the new criteria in full, says Meilan Han, a pulmonologist and researcher at the University of Michigan in Ann Arbor who is both a COPDGene investigator and a member of GOLD's scientific committee. Still, most COPD researchers are coming around to the idea that there is a group of people that research has long overlooked. "We have these symptomatic patients with a real problem that has no name, whether they have COPD or not," Han says.

The GOLD standard

COPD was first defined in the late 1950s, but it was largely neglected by researchers until the 1990s. The attitude towards patients was, "just stop smoking", Crapo says. The only available drugs were borrowed from asthma. So in 1997, a group of pulmonology researchers, as well as representatives from the World Health Organization and the US National Heart, Lung, and Blood Institute, formed GOLD as a way to raise awareness of COPD, standardize its diagnosis and encourage research on prevention and treatment.

Spirometers were already used at the time for conditions such as asthma, and they became the tool of choice to determine whether a person's breathing was obstructed. A spirometer is essentially a set of tubes attached to sensors that measure airflow. To test for COPD, a person is told to fill their lungs and forcefully breathe into the spirometer, which measures the amount of air that is pushed out.

To determine whether a person's airways are obstructed, clinicians compare the amount of air the patient can blow out in one second, called the forced expiratory volume (FEV1), to the total volume of air that they can exhale, known as forced vital capacity (FVC). According to GOLD, a person can be diagnosed with COPD if the ratio of FEV1 to FVC is below 0.7 – meaning the person exhales less than 70% of the air in their lungs in one second.

The American College of Physicians, the US Food and Drug Administration and the European Medicines Agency have all adopted the GOLD criteria. But Crapo calls them "the golden handcuffs", because the strict cut-off for diagnosis excludes two populations of patients.

First, there are those who experience episodes of intense symptoms called exacerbations, but pass the spirometry test with flying colours. Han is leading a project, called the subpopulations and intermediate outcome measures in COPD study, which has found that this group of people have airway thickening on CT scans and that their symptoms are similar to those seen in people with first- or second-stage COPD³.

"We have these symptomatic patients with a real problem that has no name, whether they have COPD or not."

The second group left out also has symptoms, exacerbations and a low FEV1, but, for whatever reason, the total lung volume of people in this group is also low, putting their spirometry ratio above 0.7. This is referred to as preserved ratio impaired spirometry, or PRISm. Those affected are prone to symptoms such as breathlessness and coughing that can interfere with normal physical activity such as walking. They also have a higher risk of death compared with people with normal FEV1 values. People can have PRISm for a variety of reasons, but for a long time it was assumed that most had fibrotic lung disease.

The COPDGene study excluded individuals with any fibrotic lung disease. This allowed researchers to conduct a long-term, detailed comparison of the health of smokers who fell into the PRISm group with those who met the GOLD criteria or had normal spirometry. Participants had clinical examinations, spirometry tests, CT scans of their lungs and blood tests at an initial assessment and then again five years later. The goal was to find genes or clinical features that could help to predict which smokers would develop COPD and how fast it would progress.

It turned out that current spirometry-based measures used for diagnosis were not the strongest predictors of worsening disease and death, says John Hokanson, who is head of epidemiology for COPDGene, and based at the Colorado School of Public Health in Aurora. His team's analysis revealed that CT evidence of emphysema (a condition in which the air sacs of the lungs are damaged) and inflammation in

the airways were the best predictors of disease progression and mortality⁴. The more extensive the airway inflammation, emphysema or both, the more likely it was that the person's disease would progress or that they would die, regardless of spirometry values.

People with signs of emphysema tended to follow the classic trajectory of COPD: first developing a low spirometry ratio but with normal FEV1, then moving on to full-blown disease. People with CT evidence of airway inflammation, however, had a completely different disease course. Half of them already had COPD, as defined by GOLD. The other half started with PRISm and, after five years, nearly 30% had developed stage 2, 3 or 4 COPD – skipping the earliest stage that would be identified by spirometry. Importantly, the PRISm in these people was not the result of fibrosis or some other condition – an indication that the disease process that led to COPD was underway years before they received an official diagnosis.

When he first saw the data, Crapo told the epidemiology team, "Oh my gosh, you just changed the diagnosis of COPD." The researchers had revealed a substantial group of people who don't meet the current COPD definition, but are nonetheless at high risk of dying from the disease. He thinks that these people should be identified and treated as early as possible – and that the best way to do that is to create several categories of COPD defined by a combination of symptoms, CT imaging, exposure to risk factors, and a low FEV1 or FEV1:FVC ratio.

Mixed reactions

Crapo is not alone in thinking that COPD diagnosis needs a revamp. "I had no trouble finding 100 other authors to put on the paper," he says. But there are doubts about whether the COPDGene proposal is the best way forward.

Even some co-authors of the proposal stress that it needs refinement. "I don't think that our proposed diagnostic criteria is the ultimate best classification," says Edwin Silverman, a pulmonologist at Brigham and Women's Hospital in Boston, Massachusetts, and COPDGene co-director. As the COPDGene team learns more about the biology behind the patterns they're seeing, he says, its scheme will be updated.

Han says she's not convinced that the airway inflammation and emphysema pathways will encompass all people with COPD. The relationship between each pathway and mortality risk is statistically complex and is based on data from people in the United States aged 45 or older who smoked heavily – at least one pack of cigarettes per day – for at least a decade



Meilan Han (right) is investigating the different forms of chronic obstructive pulmonary disease.

and often much longer. It's unclear whether Crapo's proposed criteria would work well in other groups, including the 10–20% of people with COPD who have never smoked.

On this point, Crapo and Hokanson are encouraged by data from other long-term population studies that have included non-smokers. An analysis of a population study that included nearly 5,500 smokers and non-smokers aged 45 and over in the Netherlands showed that half of people with PRISm progressed to COPD within four-and-a-half years⁵. "With respect to PRISm, we entirely replicate [the COPDGene] findings," says lead author Guy Brusselle, a respiratory physician at Ghent University Hospital in Belgium. His team is now analysing CT images from a subset of the participants of the Dutch study to see whether it can also replicate COPDGene's findings on the airway inflammation and emphysema disease pathways.

Meanwhile, Hokanson's team is analysing the third wave of COPDGene cohort data, and is finding that ten years after the start of the study, airway inflammation and emphysema are still strong predictors of disease progression and mortality. The team has also found that two genetic signatures linked to COPD align neatly with the two disease pathways. For Hokanson, that is strong evidence that these are real biological processes that lead to COPD, but he acknowledges that there are still a lot of gaps to fill.

Some critics argue that COPDGene's proposal is just not practical. Franssen says that the reliance on CT imaging makes it infeasible outside high-income countries. "It really

conflicts with the basic idea of GOLD, that it should be simple and applicable all over the world," he says. However, others argue that CT imaging is becoming more widespread, especially as part of lung-cancer screening programmes.

"We'd like to know how best to treat them, but without any evidence we can't make recommendations."

Brusselle sees considerable benefits to drug development that could come from expanding the technology's use in diagnosis. Just sorting people into two general groups of airway-inflammation-dominant or emphysema-dominant COPD would mean more-focused clinical trials, which are much needed in a field plagued by failure. As a clinician, however, he doesn't think that the COPDGene scheme offers much for patient care. It's based on statistical risk, and includes eight classifications such as possible or probable COPD. "You can't tell a patient, 'you have probable COPD,'" Brusselle says. "We need other terms."

Evidence gap

Crapo had planned to argue for revising the diagnostic criteria at a meeting of the American Thoracic Society in May. However, the meeting was cancelled as a result of the COVID-19 pandemic, and it is currently unclear when issues such as these will be discussed.

Han has already briefed the GOLD scientific committee on the COPDGene data at the European Respiratory Society meeting last September, and she suspects that it will look for formal ways to define the groups of patients who don't meet the spirometry criteria but who are at risk of COPD or have COPD-like symptoms.

David Halpin, a consultant physician at the Royal Devon and Exeter Hospital, UK, who serves on GOLD's scientific committee and board of directors, says he doesn't think there are enough data about these patients to assign formal diagnoses – especially because GOLD can't make evidence-based-treatment recommendations for them. "We'd like to know how best to treat them, but without any evidence we can't make recommendations," he says.

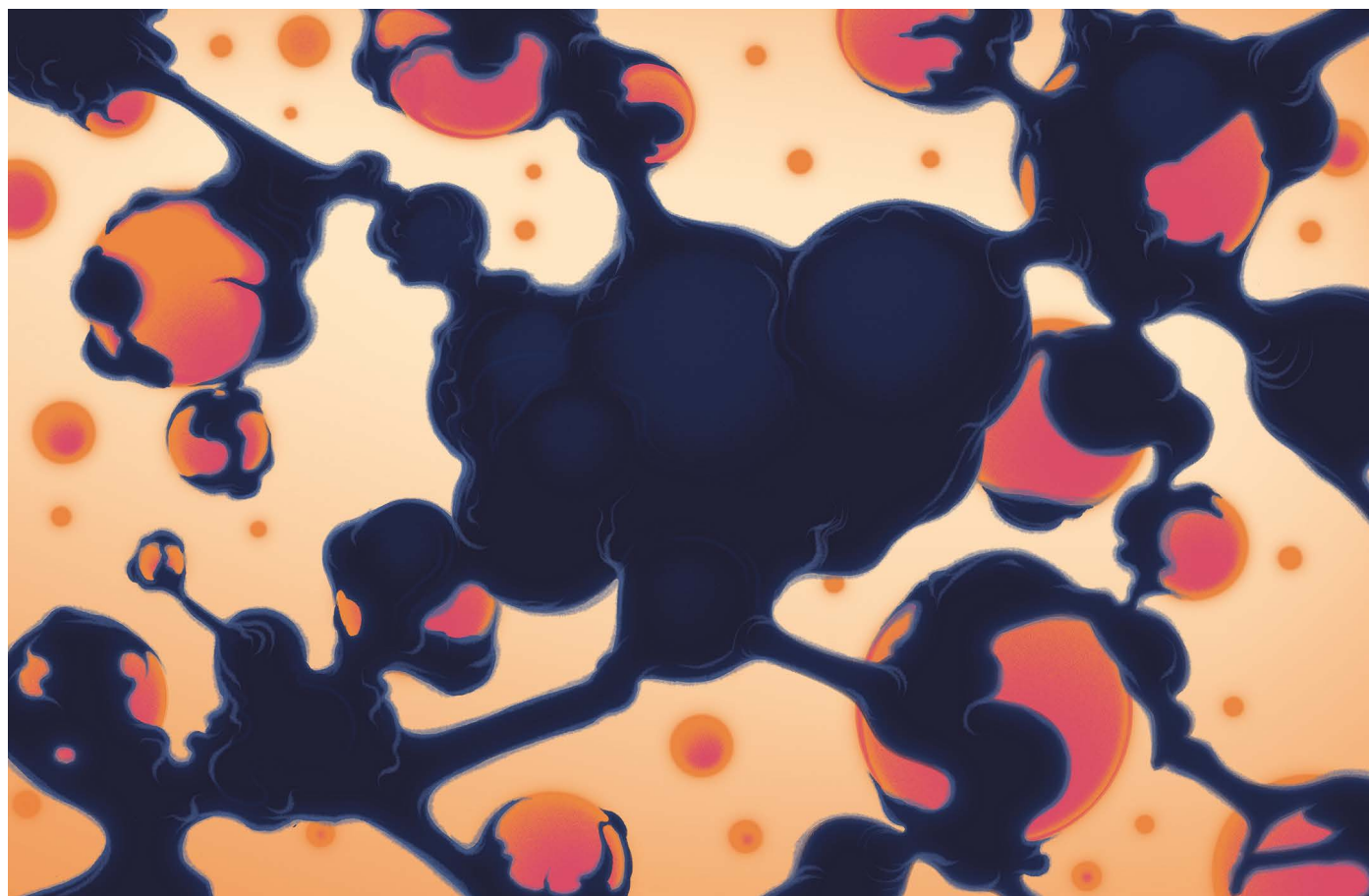
Han says this puts GOLD in a catch-22 situation: the organization can't recommend treatments for these patients without clinical trial evidence, but without names for these conditions there are no regulatory frameworks for such trials to take place, and drug companies are hesitant to enter the space. To help fill the evidence gap, Han and her colleagues are recruiting symptomatic patients with normal spirometry results to test whether a combination of two bronchodilators – medication that relaxes lung muscle and widens the airways – reduces their symptoms and improves their quality of life. There are no drug trials in the works for people with PRISm.

Crapo says that people with PRISm in the COPDGene cohort who happen to be receiving treatment tend to score higher on quality-of-life scales, but the numbers are small and the study is not designed to test interventions. He hopes that his proposal will encourage pharmaceutical companies to start studying these patients more systematically, and has been meeting with industry researchers to offer advice on designing such trials.

Crapo knows it's unlikely that GOLD will change the diagnostic criteria for COPD immediately, if at all. And he is aware that the proposed criteria need refinement and further study. But he firmly believes that waiting for lung function to decline before making a diagnosis is waiting too long. "Every single PRISm patient has high risk for progression and mortality," he says. "That's got to be recognized."

Amanda Keener is a science writer in Littleton, Colorado.

1. Young, K. A. et al. *Chronic Obstr. Pulm. Dis.* **6**, 414–429 (2019).
2. Lowe, K. E. et al. *Chronic Obstr. Pulm. Dis.* **6**, 384–399 (2019).
3. Woodruff, P. G. et al. *N. Engl. J. Med.* **374**, 1811–1821 (2016).
4. Kinney, G. L. et al. *Am. J. Epidemiol.* **187**, 2109–2116 (2018).
5. Wijnant, S. R. A. et al. *Eur. Respir. J.* **55**, 1901217 (2020).



SAM CHIVERS

Zombies in the lungs

The role of senescent cells in chronic obstructive pulmonary disease is beginning to be unpicked. **By Anthony King**

A senescent cell is like a classic movie monster – it exists in an ‘undead’ state. It no longer divides, but it is resistant to death. It is also super-sized, and produces an unusual number of proteins, many of which stoke inflammation. Almost all cells have the capacity to enter this zombie-like state, given the right circumstances. Radiation exposure, too much or not enough oxygen and certain toxins, such as those found in cigarette smoke, can all provide the spark, says James Kirkland at Mayo Clinic in Rochester, Minnesota, who is a leader in the field of cellular senescence.

Unlike the zombie armies of Hollywood films, senescent cells can be useful. When insults to DNA threaten to push a cell into a cancerous state, senescence can come to the rescue. The potential cancer cell is lulled into

a zombie-like state in which it cannot divide, and therefore cannot be cancerous. Senescent cells also draw elements of the immune system, including macrophages and natural killer cells, to their location. This helps to clear up cellular debris and any toxins that might have pushed the cells to become senescent in the first place. Senescence is also involved in wound repair, and even the initiation of childbirth.

But it is not all good news. There is a growing suspicion that senescent cells also have a leading role in triggering age-associated diseases, including chronic obstructive pulmonary disease (COPD). Put simply, researchers are beginning to worry about the zombies in our lungs.

Both of the major risk factors for the disease – ageing and smoking – are known to bring

about senescence in lung cells. Evidence also suggests that a high burden of senescent cells in the lungs is involved in the development of some of the features of COPD, such as inflammation and emphysema (damage to air sacs in the lung), although researchers lack the tools to be certain that this is a cause.

Just five years ago, few researchers would associate these cells with COPD – the third leading cause of death worldwide. Now, interest in the role of senescence in chronic lung disease is growing, potentially leading to new treatments.

Biological links

The hallmarks of COPD are emphysema and inflammation – a process involved in the hardening or fibrosis of the airways, causing them to become obstructed. Together,

STOPPING THE SPREAD OF SENESCENCE

Some researchers think that senescence might spread between cells and tissues. Could this explain why people with COPD are also likely to have other conditions?

People with chronic obstructive pulmonary disease (COPD) live with more than just airway obstruction. “One of the features of COPD is that it is almost always associated with other diseases of accelerated ageing, particularly cardiovascular disease,” says respiratory scientist Peter Barnes at the National Heart and Lung Institute in London. Cardiovascular disease is both more common and more likely to be a cause of death in people with COPD. Hormonal, metabolic, psychiatric and neurological disorders, as well as gastrointestinal disease, are also more common in people with COPD⁶. And some researchers think that cellular senescence might help to explain the high rate of comorbidities associated with COPD.

Senescent cells have the ability to convert other cells to senescence — even at a distance. “Senescence spreads from cell to cell,” says Kirkland. “You can find a predominance of senescent cells at the site of a lung disease, but in those individuals, you find senescence cells elsewhere too.”

When Kirkland’s group transplanted a relatively small number of senescent cells into young mice, this was enough to cause persistent physical problems for the animals. In older mice, even fewer cells were required to cause a problem. Eliminating the cells alleviated physical dysfunction and helped the mice to live longer⁷.

The lungs are well-vascularized organs that are regularly exposed to pollutants in air and can become overwhelmed by repeated toxic insults, which can lead to the development of numerous senescent cells. Some researchers think that this senescence might spread to other tissues.

Lee has begun to look for immune cells in bone marrow and blood in genetically altered mouse models to see whether smoking- and senescence-mediated COPD extends beyond the lung. “Many of our COPD patients have significant skeletal muscle atrophy and mitochondrial abnormalities in muscle cells. So this goes beyond the lung,” she says.

these phenomena cause shortness of breath, wheezing, a chronic cough and lack of energy.

COPD is usually diagnosed in people over the age of 40, and around half of all cases can be attributed to smoking. The habit is also known to promote cellular senescence. “Cigarette smoking is an oxidative stress to cells in the lung. That is the mechanism that puts them into senescence,” says Peter Barnes, a respiratory scientist at the National Heart and Lung Institute in London. The association has led some to suspect that senescence plays a part in triggering the symptoms of COPD. “We know that senescent cells can produce a low-grade inflammatory response, which is identical to what we see in COPD,” says Barnes. And lung biopsies taken from people with mild to moderate COPD also show signs of senescent cells, suggesting that these cells could be a cause, rather than a consequence, of COPD.

Left unchecked, the cells will secrete molecules that promote inflammation and bring about degradation of extracellular matrices — behaviour that is referred to as the senescence-associated secretory phenotype, or SASP. A build-up of senescent cells in the lungs also seems to limit the potential for tissue renewal. “Normally, these senescent cells get cleared, and this allows tissue architecture to be maintained,” says Victor Thannickal, a respiratory scientist at the University of Alabama at Birmingham. But in people with COPD, he thinks, clearance cannot keep pace with the cells’ creation. “When they don’t get cleared, then the accumulation of senescent cells can cause harm,” he says.

“When they don’t get cleared, then the accumulation of senescent cells can cause harm.”

Older people seem to be more susceptible to chronic illness caused by cellular senescence. This could be because the immune system deteriorates with age, which might impede the removal of senescent cells. Constant stimulation by toxins from cigarette smoke and pollution accelerates senescence and might exhaust the immune system, resulting in a slowing down of the body’s ability to deal with it. Eventually, a threshold might be crossed “beyond which the lung is incapable of clearing senescence cells”, says Thannickal. The higher numbers of cells can then ratchet up pro-inflammatory and pro-senescent secretions, causing tissue disruption.

But not everyone is on board with linking senescence to both the symptoms and the

causes of COPD. Cellular senescence has not yet been proved to be the prime driver of COPD. “We know that smoking triggers COPD and senescence, but perhaps senescence is a by-product of ageing or a by-product of COPD,” says Irfan Rahman, a biochemist at the University of Rochester Medical Centre in New York. In 2018, he reported that although ageing and cigarette smoking caused senescence in the lungs of mouse models of COPD, this did not worsen the severity of their symptoms¹. “We were unable to find a link between senescence and COPD in response to tobacco smoke, at least in mice,” he says.

Other researchers, however, think that cellular senescence will turn out to be a crucial driver of COPD. When Patty Lee, a pulmonologist at Duke University School of Medicine in Durham, North Carolina, interrupted early expression of senescence genes in mouse models of COPD, she found that these rodents did not go on to develop emphysema².

Search for signs

One reason for researchers’ uncertainty is that senescent cells in the body can be difficult to distinguish from healthy cells. “It’s a mess,” says Judith Campisi, a cell biologist at the Buck Institute for Research on Aging in Novato, California. “We now have maybe a dozen different biomarkers of senescent cells, but the problem is that no single biomarker is exclusive to senescent cells.” Reliable methods of detection and tracking are needed to fully understand the role of senescent cells in COPD, including the possibility that they are involved in the development of other conditions that commonly occur alongside COPD (see ‘Stopping the spread of senescence’).

Kirkland’s group is working on blood, urine and epigenetic tests so that researchers can get a handle on the burden of senescent cells in each patient. As well as certain secretory factors, the researchers want to monitor exosomes — small vesicles that senescent cells secrete in large quantities (see page S10). “With exosomes, you can tell which cell type shed them, and then you can look at their cargo for markers of senescence,” says Kirkland. Others, such as Lee, think advances in imaging will make it possible to see inside cells in the lung to detect and quantify their senescence.

Treatment dreams

Although much about the role of senescent cells in COPD is not understood, some researchers are already contemplating targeting senescence to treat COPD and other age-associated conditions.

Treatments that aim to stymie inflammation in the lungs, such as cytokine blockers, have



Judith Campisi (centre) and her colleagues are testing drugs that destroy senescent cells.

not yet proved effective. “The thinking was that inflammation drives the pathological changes in COPD, and a way of stopping the disease is to reduce this inflammation,” says Barnes. But disappointing results led to speculation that it might be better to deal with the source of inflammation – senescent cells.

The first approach to tackling senescence is a class of drugs called senostatics, which block the molecular pathways that lead to senescence. The best-studied example is the mTOR pathway, a molecular system involved in cell growth and survival, and in protein manufacturing. In 2018, researchers showed that activation of the mTOR pathway in lung vascular cells or alveolar epithelial cells of mice prompted senescence in the lungs and caused COPD-like problems³. Barnes suggests that inhibiting the mTOR pathway could be a valuable therapeutic approach.

Rapamycin, a natural compound that inhibits mTOR, has been shown to increase lifespan in mice, possibly by putting a dampener on senescent secretions. Similarly, the widely prescribed diabetes drug metformin, which increases production of the mTOR inhibitor AMPK, reduces emphysema and inflammation in mice⁴. “We don’t have evidence in humans, but we think these studies are feasible,” says Barnes. A trial is under way to

see whether metformin can reduce age-related disease in people in the United States.

A second potential approach is to destroy senescent cells. But this is no mean feat. In the lab, senescent cells survive conditions that easily kill normal cells. They also secrete compounds that act as a shield against their own killer secretions. This “allows them to survive, while killing everything around them”, says Kirkland. To defeat them, researchers are

“We were unable to find a link between senescence and COPD in response to tobacco smoke.”

looking to drugs that can knock out these cells’ shields. “Senolytics work by transiently disabling those pathways, for just a few minutes, and allow the senescent cells to commit suicide,” Kirkland explains. Unity Biotechnology, a start-up company in Brisbane, California, has had promising results from an initial phase I trial to treat osteoarthritis of the knee with its senolytic, UBX0101. Calico Life Sciences in San Francisco, California, a biotech company backed by Google, is also eyeing senolytics.

In a 2019 study, Kirkland tested a

combination of two senolytic agents in 14 people with idiopathic pulmonary fibrosis, a respiratory condition involving irreversible scarring of the lungs⁵. The compounds – a cancer drug called dasatinib, and a compound found in many fruits and vegetables called quercetin – triggered the death of senescent cells and improved the participants’ physical function. A phase II study is under way. “The big question is, will it work for COPD,” says Campisi, who co-founded Unity Biotechnology. “We are working on that now.”

Hope and hype

Some researchers are wary, however, about getting too carried away. Removing every senescent cell in the lungs might have damaging side effects. “Which cell types senolytics target is going to be important,” warns Lee. Safety concerns partly explain the initial focus on idiopathic pulmonary fibrosis – on average, people die within four years of diagnosis. “Senolytics should start off with very serious life-threatening conditions for which there is no good treatment,” says Kirkland. “We don’t know the side effects of these drugs yet.”

Some fear that research on the potential clinical relevance of senescent cells could be overwhelmed by hype. One COPD researcher, who wanted to remain anonymous, said they had never heard of senescence five years ago, but it now seems almost obligatory to mention the phenomenon in grants or papers on COPD. Thannickal similarly notes that what was a trickle of reports on the topic five to ten years ago has turned into a waterfall. But Barnes argues that there is good reason to pay attention to senescence. After all, COPD affects one in ten people over the age 40, and there is an urgent need for treatments that do more than just manage symptoms. “It is such a common disease,” he says. “It is really good to test out some of these ideas.”

The zombie hordes of films and books are usually defeated. And although it is too early to tell whether COPD and other age-associated diseases will follow the same script, zombie cells are at least now firmly in researchers’ cross hairs. “There’s recognition now of the importance of senescence and lung ageing in COPD,” says Lee. “It is certainly on our target list,” agrees Kirkland. “We and other labs here are working around the clock.”

Anthony King is a science writer in Dublin.

1. Rashid, K. *et al. Sci. Rep.* **8**, 9023 (2018).
2. Kim, S.-J. *et al. Aging Cell* **18**, e12914 (2019).
3. Houssaini, A. *et al. JCI Insight* **3**, e93203 (2018).
4. Cheng, X.-Y. *et al. Oncotarget* **8**, 22513–22523 (2017).
5. Justice, J. N. *et al. EBioMedicine* **40**, 554–563 (2019).
6. Yin, H.-L. *et al. Medicine* **96**, e6836 (2017).
7. Xu, M. *et al. Nature Med.* **24**, 1246–1256 (2018).



Irfan Rahman exposes cells to smoke as part of research into exosomes and COPD.

Care packages

Vesicles released in response to cigarette smoke might trigger COPD, but engineered versions offer possible therapy. By Jyoti Madhusoodanan

In the 1980s, researchers found that healthy cells release small, membrane-wrapped packages that are now known as exosomes. They originate deep inside cells, where they are loaded with cargo including specific proteins and RNA before being released to travel beyond the cell.

Initially, researchers thought of exosomes as a means of intercellular communication. “At the time, people thought exosomes were only released to relay neurotransmitters or hormones,” says pulmonologist Yang Jin of Boston University, Massachusetts. “Their importance has only been recognized in the last ten years or so.”

Now, scientists know that nearly all cells shed exosomes. And Jin and others have found that these vesicles might be key to the symptoms of chronic obstructive pulmonary disease (COPD).

People with COPD – one of the leading causes of death worldwide – experience wheezing, fatigue and chronic coughing. It is especially prevalent in smokers, and research

has found that both smokers and people with COPD have an increased number of exosomes circulating in their blood. The contents of these vesicles also differ markedly from those seen in non-smokers without the disease. “We don’t know the true triggers of COPD,” Jin says. “Looking at the cargo of vesicles in different groups of patients could potentially hold answers about how this disease develops.”

In addition to working out the role of exosomes in the development of disease, several researchers are eyeing their therapeutic potential. Early studies suggest that vesicles derived from stem cells can aid tissue repair, and some scientists are considering the possibility of engineering vesicles to carry drugs to diseased tissues. But these efforts have been held back by a dearth of standardized methods to isolate and study vesicles. Advances in techniques over the past few years – and greater scientific consensus in creating standards for research into extracellular vesicles – are pushing the field forward.

Some of the clearest evidence linking exosomes to the symptoms of COPD emerged in 2019. While trying to understand how a particular protein exited immune cells, Edwin Blalock, a pulmonologist at the University of Alabama at Birmingham, found it inside exosomes, along with an unexpected travelling companion: the enzyme neutrophil elastase¹.

Elastase is a prominent player in COPD. The enzyme wears down the stretchy fibres of elastin and collagen that keep the lungs flexible. In healthy individuals, cells counter elastase’s effects with an anti-protease called α 1-antitrypsin (α 1AT), and COPD was long considered the result of an imbalance between these two proteins. This view is bolstered by the fact that people with a genetic deficiency in α 1AT are at much greater risk of developing COPD – even if they have never smoked – than are non-smokers without the mutation. The idea that higher levels of neutrophil elastase are linked to COPD “has been a cornerstone of the study of COPD for over six decades”, says Blalock. “But the levels of elastase typically seen were never high enough to counter α 1AT activity. That was the conundrum.”

Blalock and his colleagues found that when elastase was packed on the surface of exosomes, it was protected from neutralization by α 1AT. These exosomes also bore a marker called Mac-1 that helped them to bind to the extracellular matrix, where elastase then digests matrix fibres. The loss of elastin and collagen from the extracellular matrix causes lung tissue to become less flexible and alveolar spaces to widen, which in turn reduces the efficiency with which the lungs transfer oxygen and carbon dioxide into and out of the body.

When exosomes from people with COPD were injected into mice, the animals developed signs of COPD, including emphysema¹. “This is the first instance of being able to have exosomes transfer a disease phenotype from a human to a mouse,” Blalock says. “It’s surprising, especially the rapidity with which the mice developed COPD after they first encountered these exosomes, and I think it points to their potency as effectors of damage.”

Spurring symptoms

Neutrophils are not the only source of exosomes implicated in COPD. In healthy people, lung epithelial cells usually release exosomes containing a protein called CCN1. But Jin’s team found that when mice were exposed to cigarette smoke – about the equivalent of around 70 cigarettes a day for 3 months – lung epithelial cells instead released a fragmented form of the protein directly into bronchial fluids². The intact

form of the protein inside exosomes modulated inflammatory proteins in the lung and helped to maintain homeostasis after exposure to cigarette smoke. But the CCN1 fragments not encapsulated in vesicles caused a spike in the production of two proteins that digest the extracellular matrix, causing cells and tissues to die. The reason, Jin suggests, is that smoking and other stressors alter how proteins such as CCN1 are tagged for processing, resulting in the production of abnormal fragments that are not wrapped in an exosome.

Jin and others are also looking at microRNAs in exosomes; these are more stable and easier to detect than proteins. Several microRNAs are enriched in extracellular vesicles from lung epithelial cells exposed to cigarette smoke, according to one study³. Researchers found that one of these, miR-210, reduced autophagy, a process that is essential to clearing away damaged cells. The microRNA also increased the formation of collagen and cells associated with fibrosis, which stiffens lungs. All these functions could contribute to the development of COPD, says Takahiro Ochiya who studies exosomes at Tokyo Medical University, lead author of the study.

Because exosomes carry multiple molecules, it has long been hoped that their contents could be used as diagnostic or prognostic biomarkers. Not all those who smoke develop COPD, and not all those who have COPD are smokers. The contents of extracellular vesicles might help to “figure out whether a person has the potential to develop emphysema or not”, Jin says.

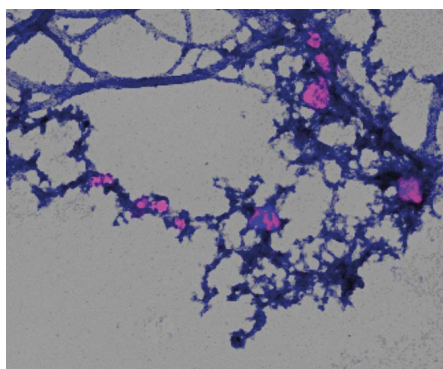
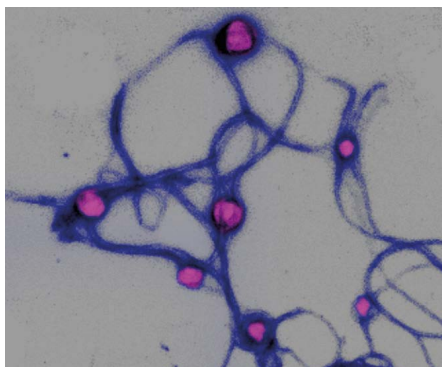
Blalock suggests that future studies of exosomes from activated neutrophils should

“If we keep an open mind, there may be multiple cell types or vesicle types we could use for therapeutics.”

examine whether these vesicles occur in all smokers, or just a sub-population who might be at greater risk of COPD as a result. “If it’s only a sub-population of otherwise healthy smokers, are they the ones to go on to develop COPD?” he says. “If that’s the case, we would have a biomarker to identify the people who smoke who are likely to get the disease.”

Microscopic mules

Knowing the molecular triggers that exosomes carry is also a step towards finding drug targets and designing better therapies. Elastase is one possibility. Because the enzyme is shielded from its natural inhibitor while attached to exosomes from activated neutrophils, it is



Elastase on neutrophil-derived exosomes (pink) breaks down collagen (blue).

possible that an intervention that dissociates elastase from the exosome could make the enzyme susceptible to a person’s α_1 AT once again, Blalock suggests.

Engineered vesicles could also be used to carry drugs to specific sites of tissue damage. “Current therapies for COPD are just analgesic or palliative,” says Irfan Rahman, who studies environmental medicine at the University of Rochester in New York. “We give steroids, β -agonists or bronchodilators just to open up the lungs, but the destruction continues.”

Last year, Rahman and his colleagues reported that, in mice, vesicles derived from mesenchymal stem cells protected lung tissues from the damage caused by exposure to cigarette smoke⁴. And in ongoing studies, Ochiya and his colleagues are evaluating whether a spray delivered directly into the trachea, containing vesicles harvested from healthy lung cells, can reverse the damage caused by COPD.

Jin’s team is taking a different approach. Instead of using vesicles derived from healthy cells, it is aiming to manipulate the contents of vesicles to deliver drugs, proteins or microRNAs to treat the symptoms of lung disease. Because vesicles share the surface markers of the cells they are derived from, they can be directed specifically to diseased tissues. “This decreases a lot of side effects that are caused by medications affecting

non-target tissues,” Jin says.

These and other exosome-based therapies for a variety of conditions, including cancer and Alzheimer’s disease, are still in preclinical development – numerous experimental and regulatory hurdles remain.

Better standards

Most vesicle-based therapeutic strategies in COPD currently rely on vesicles released from cultured cells. But these vesicles vary widely in their contents and how they’re formed, making it tough to isolate a pure sample of exosomes and to standardize therapeutic effectiveness. “We still have no gold-standard method to harvest vesicles,” Ochiya says.

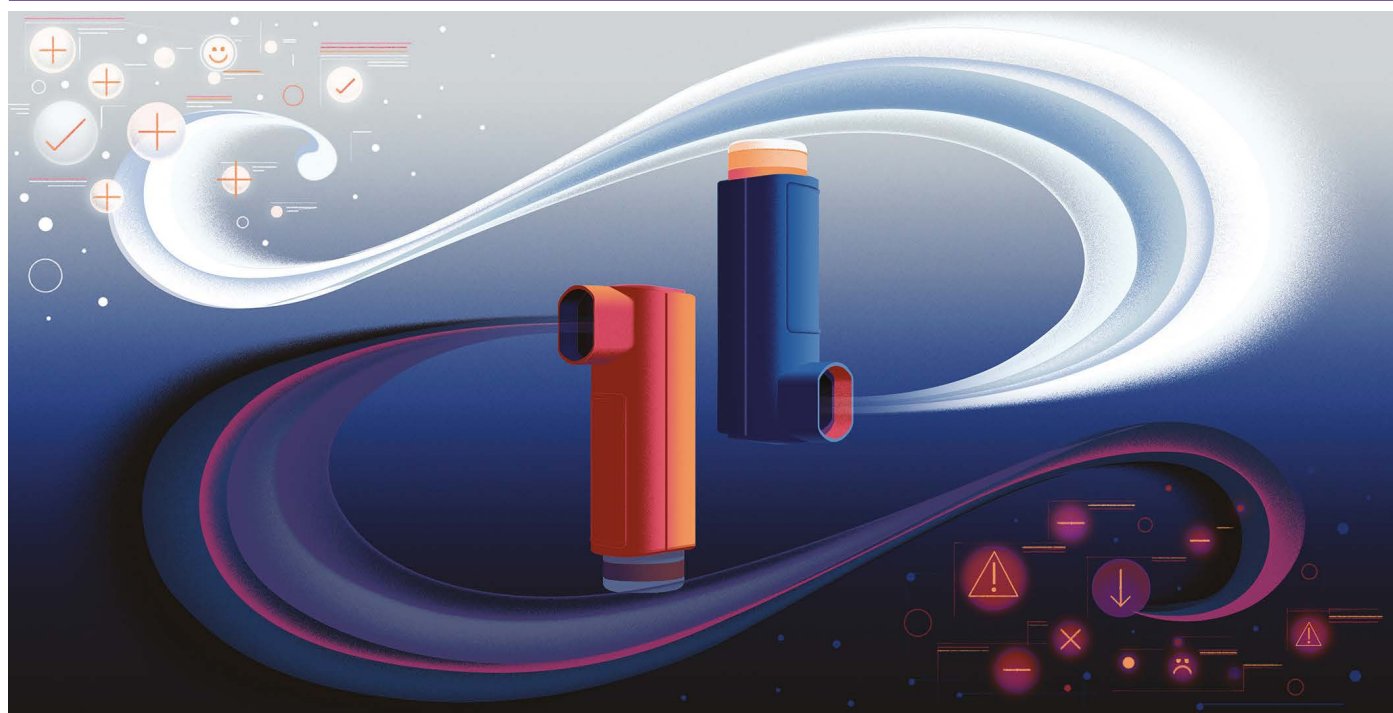
Exosomes are the only vesicles known to be produced by a cell’s internal membranes. One approach to purifying samples of vesicles down to exosomes has been to look for signs of a vesicle having passed through this production pipeline. However, focusing on how the vesicle is formed, rather than its function, might stymie efforts to develop therapeutics, says molecular biologist Kenneth Witwer of Johns Hopkins University in Baltimore, Maryland. Focusing on function could expand the range of potential therapies beyond stem-cell derived exosomes. “If we keep an open mind, there may be multiple cell types or vesicle types we could use for therapeutics,” he says. Witwer is one of a number of researchers to propose methods for characterizing vesicles that shift the focus from how they are made to their size, cargo or function.

Demonstrating that a particular batch of vesicles have uniform physical features, and then showing the vesicles’ potency in a functional assay, “would help regulators assess whether a vesicle-based product is essentially the same from one batch to the next”, Witwer says.

This standardization is crucial if exosome-based therapies are to become a reality. At present, people with COPD are treated with bronchodilators or other drugs that stave off symptoms, but do little to halt the underlying tissue damage. Therapies that rely on exosomes derived from stem cells could perform better than stem cells themselves, particularly because “exosomes may be able to go places where a cell can’t”, Blalock says. “There may be therapeutic niches that can only be accessed via exosomes.”

Jyoti Madhusoodanan is a science writer in Portland, Oregon.

1. Genschmer, K. R. *et al.* *Cell* **176**, 113–126 (2019).
2. Moon, H. G. *et al.* *Am. J. Physiol. Lung Cell Mol. Physiol.* **307**, L326–L337 (2014).
3. Fujita, Y. *et al.* *J. Extracell. Vesicles* **4**, 28388 (2015).
4. Maremmanda, K. P., Sundar, I. K. & Rahman, I. *Toxicol. Appl. Pharmacol.* **385**, 114788 (2019).



The steroid debate

Should physicians still be prescribing steroid inhalers as a first-line treatment? By Julianna Photopoulos

The first inhaled corticosteroids were developed in 1972 for use in people with asthma. The drugs, which tackle inflammation in the airways, were revolutionary in reducing the number of hospital admissions and deaths due to asthma.

It is little wonder, then, that physicians, faced with a paucity of treatment options for chronic obstructive pulmonary disease (COPD), would adopt the anti-inflammatory steroids that had proved so effective in managing asthma. The drugs became a common prescription – one of the first clinical trials to test inhaled steroids for COPD found that more than half of the people recruited between 1992 and 1995 were already receiving them¹. Now, many estimates put the use at around 75%.

And yet the evidence underpinning the efficacy of steroids for COPD is surprisingly inconclusive. “Inhaled steroids have turned out to have very little clinical benefit in COPD,” says Peter Barnes, a respiratory scientist at the National Heart and Lung Institute in London. Although the drugs might be beneficial to some people with COPD, many researchers think that the upsides are often outweighed

by the risk of side effects. As that point of view has become more widespread, guidelines on prescribing inhaled steroids for COPD are changing. The aim now is to give the drugs only to those who stand to benefit.

Balance of evidence

Inhaled steroids are commonly prescribed in combination with drugs known as bronchodilators. The main examples used for COPD are long-acting β_2 -agonists (LABAs) that widen airways by causing lung muscles to relax, and long-acting muscarinic antagonists (LAMAs) that prevent nerves from releasing chemicals that cause the airways to tighten. Used daily, bronchodilators help to manage symptoms, improve lung function and – through processes that are not well understood – prevent flare-ups, known as exacerbations.

For asthma, LABAs can be given only alongside inhaled steroids, says Leonardo Fabbrì, a respiratory researcher at the University of Ferrara in Italy – on their own, the bronchodilators increase the risk of a life-threatening asthma attack. LABA and steroid combination inhalers are also commonly used to manage

COPD, but including a steroid, rather than a combination of LAMA and LABA bronchodilators, has been called into question.

The efficacy of the two drug combinations has been extensively tested, and the findings are conflicting. In 2016, the FLAME trial, involving around 3,000 people, reported an 11% lower rate of COPD flare-ups when people used a combination of LAMA and LABA for a year, than when they used a LABA and steroid inhaler². But in 2018, the larger IMPACT study, which involved more than 10,000 people with moderate-to-severe COPD, found the opposite – a combination of LABA and steroid was associated with fewer exacerbations³.

The apparent disagreement could be because steroids work better for some forms of COPD than for others, says Daiana Stolz, a respiratory researcher at University Hospital Basel in Switzerland. Although the FLAME study suggested that steroid inhalers were outperformed by LABA and LAMA treatment, Stolz says that people who had previously experienced frequent exacerbations did respond positively to inhaled steroids. And a clinical practice study led by pharmacoepidemiologist Samy Suissa at McGill University in Montreal, Canada, found that roughly 10% of participants benefited from using inhaled steroids rather than dual bronchodilators⁴.

Suissa's study also found that people with COPD who started on LABA and inhaled steroid treatment were more likely to develop pneumonia than were those who did not receive a steroid treatment. This is a common safety concern associated with inhaled steroids.

People with COPD are already more prone to developing pneumonia than are healthy people, Fabbri says. But several studies have suggested that inhaled steroids increase the risk – nearly doubling it in some cases⁵.

New recommendations

Although there is still some debate as to who can benefit from inhaled steroids, researchers agree that the drugs have long been over-prescribed. “It’s important that we only give drugs to patients who are likely to benefit from them,” says James Chalmers, a respiratory researcher at the University of Dundee, UK.

Since 2001, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) has published a strategy for diagnosing and managing COPD, which is often used as the basis for national and regional guidelines. For many years, GOLD recommended that inhaled steroids be broadly prescribed for people with frequent exacerbations and severe loss of lung function. But over time, that advice has changed. In 2011, LAMA bronchodilators were recommended over combinations of LABA and inhaled steroids, following research that showed that the two had a similar effect on the rate of flare-ups. And in 2017, following the FLAME findings, GOLD recommended that LABA and steroid inhalers be given only when LAMA and LABA therapy fails to control symptoms.

In 2019, GOLD highlighted a biomarker – blood eosinophil count – that could be used to identify which people experiencing frequent exacerbations are most likely to benefit from inhaled steroids. Eosinophils are white blood cells that fight infection, and can contribute to airway inflammation. Ian Pavord, an airway-disease researcher at the University of Oxford, UK, found that the higher the eosinophil levels in people with COPD, the more effective steroids are at managing exacerbations⁶. In people with low eosinophil counts, steroids had little effect. In a later study, he showed that people with low cell counts are also at greater risk of pneumonia⁷.

A person’s eosinophil count can vary – levels are higher in the morning than they are in the evening, says Stolz, suggesting that multiple tests might be needed to ensure physicians have an accurate picture of their patients. But even so, the measure has turned out to be a “surprisingly good indicator” of whether people with severe COPD will respond to steroids, says Barnes. GOLD now recommends that a blood eosinophil count of more than 300 cells per microlitre is a sign that people with frequent exacerbations and severe symptoms will benefit from inhaled steroids. If a person’s eosinophil count is under 100 cells per

microlitre, inhaled steroids are discouraged owing to lack of efficacy and the increased risk of pneumonia, even if the person is experiencing frequent exacerbations.

Disparate worlds

In 2017, Chalmers and his colleagues estimated that more than 60% of people in the United Kingdom with COPD were receiving steroids as a first-line treatment⁸. Pavord hopes that including blood eosinophil count as a biomarker in the GOLD recommendations will lead to inhaled steroids being prescribed more selectively (only around 10–20% of people with COPD have eosinophil counts greater than 300 cells per microlitre). But clinical practice does not always follow GOLD recommendations to the letter. One 2019 study found that many Europeans at low risk of COPD exacerbations were still being prescribed inhaled steroids⁹.

“It’s important that we only give drugs to patients who are likely to benefit from them.”

What happens in research labs and what is done in clinical practice are different things, says Suissa. “These are two completely disparate worlds.” In some countries, he says, long-acting bronchodilators are either available only with an inhaled steroid – as would be required for asthma – or can be prescribed without a steroid only by specialists. Until this year, primary-care physicians in Israel gave patients combined LABA and steroid inhalers because they could not prescribe LABA alone, he explains.

Some researchers are also concerned that the ready availability of triple-combination inhalers that contain both bronchodilators and a steroid might lead to more people receiving steroids.

The IMPACT study found that the rate of flare-ups in people using triple therapy was 25% lower than in those using LAMA and LABA combination inhalers. The rate of pneumonia, however, was 50% higher. Several other trials have also reported lower rates of exacerbations associated with triple therapy than with dual-bronchodilator therapy, says Fabbri. He thinks that there are cases in which triple therapy could be beneficial as a first-line treatment, despite current guidelines, and says that most people with COPD will end up using it eventually. Barnes agrees that this is likely, albeit inappropriate in his estimation, simply because triple therapy is “the easiest way to manage COPD”.

While researchers and clinicians debate the best prescriptions for people with COPD, a thornier issue looms: what to do about the millions of people already receiving inhaled steroids. “There is clearly no point” in administering medicines that could do more harm than good, Pavord says. But, he admits, “it’s quite hard withdrawing treatment in very symptomatic patients”, which most are. Fabbri thinks that if the treatment seems to be working and there aren’t any other complications, it should be continued – even if it includes a steroid.

Take it away

Some evidence suggests that steroids can be safely withdrawn from people with COPD who are used to taking them. For example, an observational study in Japan found that older people with COPD who had the steroid component of their treatment withdrawn after a flare-up were less likely to die or be admitted to hospital than were those who stayed on the steroid¹⁰. The 2014 WISDOM study also found that gradually discontinuing inhaled steroids did not affect flare-ups in people who had been using triple therapy, although the results did suggest that continuing to use inhaled steroids was beneficial for lung function¹¹ – the importance of which has divided researchers.

As things stand, there are no international recommendations about withdrawing steroids from people with COPD. But Chalmers expects guidelines from the European Respiratory Society on who it is appropriate to withdraw inhaled steroids from, and how best to do it, to be published in May. “Hopefully it’ll start to reverse some of the overuse of steroids across Europe,” he says.

For Chalmers, it is time to move on from the inhaled-steroid debate. Even in people who do see a benefit, steroids are not very effective treatments, he argues. “We have spent too much time talking about steroids,” he says. “We need to invest more energy into finding better treatments.”

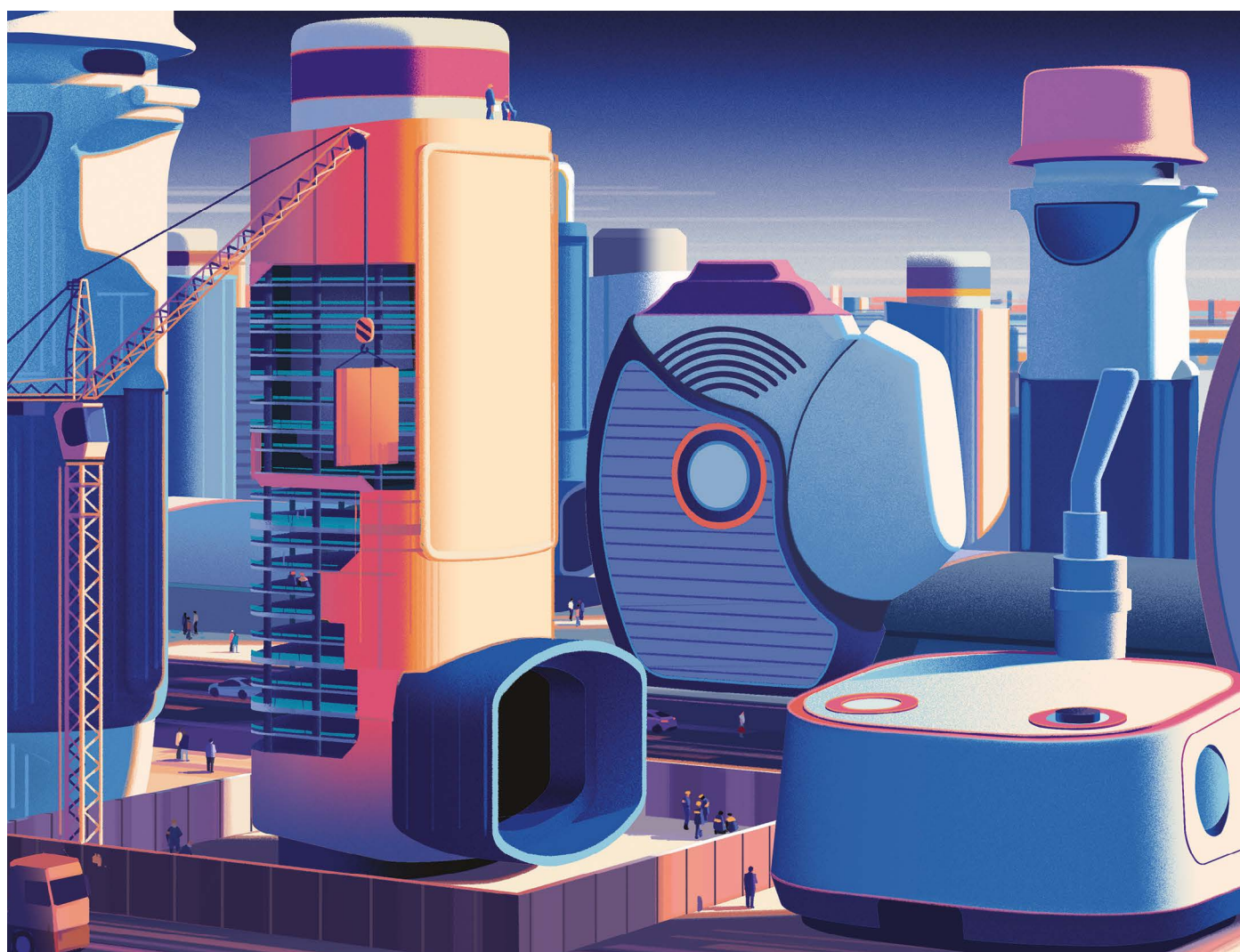
Julianna Photopoulos is a science journalist near Thessaloniki, Greece.

1. Burge, P. S. et al. *Br. Med. J.* **320**, 1297 (2000).
2. Wedzicha, J. A. et al. *N. Engl. J. Med.* **374**, 2222–2234 (2016).
3. Lipson, D. A. et al. *N. Engl. J. Med.* **378**, 1671–1680 (2018).
4. Suissa, S., Dell’Aniello, S. & Ernst, P. *Chest* **155**, 1158–1165 (2019).
5. Finney, L. et al. *Lancet Resp. Med.* **2**, 919–932 (2014).
6. Pascoe, S. et al. *Lancet Resp. Med.* **3**, 435–442 (2015).
7. Pavord, I. D. et al. *Lancet Resp. Med.* **4**, 731–741 (2016).
8. Chalmers, J. D. et al. *npj Prim. Care Resp. Med.* **27**, 43 (2017).
9. Vestbo, J. et al. *Int. J. Chron. Obstruct. Pulmon. Dis.* **14**, 853–861 (2019).
10. Jo, T. et al. *ERJ Open Res.* **6**, 000246 (2020).
11. Magnussen, H. et al. *N. Engl. J. Med.* **371**, 1285–1294 (2014).

THE INHALER MAKEOVER

Environmental concerns and the perennial problem of poor inhaler technique are driving change in the devices used to deliver respiratory drugs.

By Sarah DeWeerd



The most important advance in inhaled therapy for chronic obstructive pulmonary disease (COPD) in the past 50 years didn't come from the discovery of a molecule or biological mechanism but instead from an international treaty that phased out the use of chlorofluorocarbons (CFCs). These chemicals, which had been found to

damage the atmospheric ozone layer, were used as propellants in inhalers that deliver drugs to the lungs.

The adoption of the Montreal Protocol in September 1987 by the United Nations created a sudden need to find alternatives to CFCs, and injected a much needed spark of innovation into the inhaler field, says Stephen Stein, an aerosol scientist at manufacturing company 3M in St. Paul, Minnesota, who has

been involved in inhaler research for more than 20 years. As companies raced to develop the first non-CFC inhaler, they also “took the opportunity to improve upon a technology that really had kind of stagnated over the past decades”, Stein says.

The result was a plethora of inhalers and inhaled drugs that transformed the care of people with COPD. “We went from allowing patients to put up with their symptoms to

actively managing their symptoms,” says Jane Scullion, a consultant respiratory nurse at University Hospitals of Leicester, UK. “If you get it right for a patient, it can transform their lives.”

Now, history is repeating itself – with a twist. Although the field isn’t stagnant by any stretch, it is struggling to solve long-standing problems: getting people to use their inhalers as prescribed and with the correct technique so that the medication reaches their lungs. Meanwhile, increasing attention on the environmental impacts of hydrofluoroalkane (HFA) propellants, which replaced CFCs but are themselves powerful greenhouse gases, is spurring innovations in inhaler design.

The mother of invention

The use of inhaled therapies for respiratory diseases goes back 3,500 years to ancient Egypt. But the modern-inhaler era began in 1956, when scientists at Riker Laboratories in Minnesota (Riker was acquired by 3M in 1970) introduced the Medihaler: the first portable inhaler.

The Medihaler was designed to treat asthma (inhaled therapy for COPD didn’t begin until the early 1960s). It was a metered-dose inhaler (MDI), and would be familiar to current users of these devices. The user presses on a canister to release a puff of drug mixed with a propellant, while inhaling slowly and steadily to draw the medicine into the lungs.

When the Montreal Protocol was signed three decades after the Medihaler’s introduction, the vast majority of inhalers in use were MDIs that used CFC gases as the propellant. Pharmaceutical companies responding to the new ban found that, to accommodate new HFA propellants, they had to tweak the design of the inhalers. This gave them an opportunity to fix some of the other long-known problems of MDIs. For example, the older devices left much of the drug in the mouth and throat, so researchers made changes to allow a larger proportion of the drug to reach the lungs, such as altering the inhalers so that they release smaller particles.

Other companies took a different path, focusing on developing a type of inhaler that uses no propellant at all. These dry-powder inhalers (DPIs) had been invented in the 1850s, but the technology had hardly advanced by the time of the CFC ban more than a century later. With no propellant to dispense their contents, DPIs instead rely on a quick, deep inhalation to draw the drug into the lungs.

Since the Montreal Protocol, other devices such as soft-mist inhalers and modern nebulizers have also joined the mix (see ‘The right device for the right patient’). The device of choice varies from country to country,



The original Medihaler (top) and 3M’s digital smart inhaler.

depending on which strategy companies in a given region pursued. For example, MDIs remain the most popular inhalers in the United Kingdom, whereas DPIs are the top choice in Scandinavia.

Persistent imperfection

As inhaler technology advanced, the range of drugs available for treating COPD expanded. Longer-acting and more-effective bronchodilators to relax and widen the airways emerged, as did a greater variety of corticosteroids to control inflammation in the lungs. By 2011, there were more than 230 different drug-device combinations on the market in Europe.

Despite these innovations, there has been little improvement in the number of people who use their inhalers correctly. For example, a comprehensive analysis of 144 studies conducted between 1975 and 2014 found that people with COPD are still as likely as patients 40 years ago to use inhalers incorrectly¹. Mistakes can markedly reduce the effectiveness of

treatment of all types of device. Overall, 31% of people in the studies reviewed had poor inhaler technique.

In some ways, the profusion of drugs and devices only increases the challenge. Most people with COPD have multiple inhalers. These devices can require different breathing techniques, which people might have trouble remembering – especially if they are struggling to breathe.

Combination inhalers, which have two or three drugs in a single device, can alleviate some of the potential for error by reducing the number of inhalers a person has to manage. But most people will still have at least two: one for daily use to prevent symptoms from starting and another to provide quick relief. Some people get these mixed up – carefully scheduling doses of their quick-relief inhaler but using their preventive inhaler only when they feel short of breath, says Amber Martirosov, a pharmacist at Wayne State University in Detroit, Michigan. And when people fail to use their preventive inhaler properly, they are at greater risk of episodes of more-severe disease known as exacerbations.

In theory, there is a simple solution to these problems: better patient education. “Once you go through the basic steps with a patient, it’s really not that hard for them to use it – but you’ve got to take the time,” Martirosov says. Often, this task falls through the cracks for physicians, nurses and pharmacists alike. It’s also not a one-off job – technique can slip over time and regular reminders are needed.

Smarter devices

Even with good patient education, certain groups of people are likely to encounter trouble using specific types of inhalers. Some people might lack the strength and dexterity to push the canister of an MDI, or the ability to coordinate their breath with the puff of medicine. Those with memory or cognitive impairments might have trouble assembling a soft-mist inhaler. And those with severe COPD might not be able to breathe in with sufficient strength to activate a DPI.

Such physical limitations might be more common than initially thought. Over the past several years, scientists including Martirosov have found evidence that some people with advanced COPD lack the lung capacity to effectively use MDIs. “We found a subset of patients that we would teach, but they couldn’t ever fix their technique,” she says. “They couldn’t ever improve that manoeuvre.” Her team is now investigating whether switching to nebulizers is the answer for these patients.

But by and large, health-care professionals are not used to thinking about matching

THE RIGHT DEVICE FOR THE RIGHT PATIENT

When prescribing treatment for conditions such as chronic obstructive pulmonary disease, health-care professionals must consider not just the drug, but also which inhaler will best serve their patient. Each device has its drawbacks, and with more than 230 drug-device combinations available, it is not a trivial decision.



METERED-DOSE INHALER

This inhaler consists of a canister attached to a plastic mouthpiece. The user presses on the canister to release a puff of drug mixed with a propellant, while slowly inhaling to draw the drug into the lungs.

Advantages:

- ▲ Fast and simple to use — good for emergency treatment.
- ▲ Suitable for delivering most drugs.
- ▲ Stable in hot and humid climates.
- ▲ Less expensive than many inhalers.

Disadvantages:

- ▼ Contains environmentally damaging propellants.
- ▼ Requires strength and dexterity to release a puff of medicine.
- ▼ Might not be effective in people whose peak inspiratory force is low.
- ▼ Requires the patient to coordinate actuating the device and breathing in (using a holding chamber or spacer can help with this).



DRY-POWDER INHALER

Medication in these inhalers is stored in single-use capsules, or in a blister pack or reservoir from which individual doses are dispensed. Instead of a propellant, the devices rely on a quick, deep inhalation to draw the drug in.

Advantages:

- ▲ Suitable for those who have difficulty with coordination.
- ▲ Can be used to deliver most drugs.
- ▲ Simple to use for those with cognitive impairments.
- ▲ Many include a counter showing how many doses remain.

Disadvantages:

- ▼ The inspiratory force needed to use the inhaler can be problematic for people with advanced COPD, or those experiencing or recovering from exacerbations.
- ▼ Not suitable for use in hot or humid climates.
- ▼ Some devices require the preparation of individual doses.

SOFT-MIST INHALER

This inhaler does not contain propellant. Instead, the mechanical force of a spring turns a solution of medication into an aerosol. Twisting the base of the device compresses the spring, and pushing a button releases it. The user then inhales the aerosol through the mouthpiece.

Advantages:

- ▲ No need to coordinate breath and activation.
- ▲ Does not depend on a user's inspiratory flow rate.
- ▲ Targets small airways with extra-fine particles.
- ▲ Can be refilled.

Disadvantages:

- ▼ Complicated to put together.
- ▼ Often more costly than other devices.



NEBULIZER

These are machines that deliver medicine as a wet mist. After preparing the device with a dose of liquid medication, the patient turns the machine on and breathes in slowly through a mouthpiece or face mask until all the medicine has been delivered.



Advantages:

- ▲ Effective for people who can't breathe in with sufficient force to use inhalers.
- ▲ Good for emergencies and when higher doses of treatment are required.
- ▲ Suitable for those with cognitive or physical disabilities.

Disadvantages:

- ▼ Often bulky and inconvenient.
- ▼ Slower drug delivery — treatment takes 5–15 minutes.
- ▼ Requires regular maintenance and cleaning.

devices to the patient. One survey of physicians found that respondents overwhelmingly placed greatest importance on the choice of drug rather than device when deciding on treatment for people with COPD; only around one-third considered the choice of device to be highly important². “We’ve been trained for so long to just focus on the medicine and not the device,” says Jill Ohar, a pulmonologist at Wake Forest School of Medicine in Winston-Salem, North Carolina, who worked on the study.

“I don’t think I ever once had a lecture in my training to become a respiratory specialist on devices,” agrees Omar Usmani at Imperial College London. “What people don’t get is that the treatment is the drug with the device.”

Some pharmaceutical companies are betting that smart inhalers with embedded electronic components can help to improve inhaler technique and adherence to treatment. Digital therapeutics companies Propeller Health in Madison, Wisconsin, and Adherium in Auckland, New Zealand, have both introduced devices that can be attached to several types of inhaler to track a person’s use of medication through a mobile app.

The first stand-alone smart inhaler, the Digihaler, marketed by Teva Pharmaceutical in Petah Tikva, Israel, was approved in the United States in late 2018. The device records when a person uses the inhaler, as well as the rate at which they inhale, and sends this information to a smartphone app; patients can then review the data with their physician. A similar device, known as the Intelligent Control Inhaler, is being developed by 3M. It tracks use and tells the patient how to use it through a screen on the device. The 3M unit also troubleshoots some common mistakes when using inhalers. For example, it notifies people if they forget to shake the device before inhaling.

The Intelligent Control Inhaler is aimed at people with COPD, says Stein. “We interviewed and worked with COPD patients to make sure the system would be usable for them,” he says. For example, it is designed to require an inspiratory flow rate that is achievable for most people with COPD. But the device will also be appropriate for those with asthma; 3M is now working to commercialize it with pharmaceutical partners.

Climate re-emergence

The advent of smart devices might cut against other goals for improving inhalers, however. In some health-care systems, even those of wealthy countries such as the United States, cost can be a major barrier to accessing inhalers; the high-tech versions could be

even further out of reach. And although the Digihaler and Intelligent Control Inhaler are both refillable, the electronic component of each is designed to be used for only about a year. “It’s hard to fathom how that could possibly be sustainable in the long term,” says Alex Wilkinson, a respiratory physician at East and North Hertfordshire NHS Trust in Stevenage, UK. Even in the United Kingdom, which has a widespread inhaler recycling programme, less than 1% of devices are actually recycled – and disposable electronics will further add to the waste stream.



WE’VE BEEN TRAINED FOR SO LONG TO JUST FOCUS ON THE MEDICINE AND NOT THE DEVICE.”

An even more pressing concern is the impact of inhalers on the climate, particularly MDIs. It’s long been known that the switch from CFC to HFA propellants didn’t solve all the environmental problems associated with the devices. Adisa Azapagic, a sustainability analyst at the University of Manchester, UK, says that one puff from a typical MDI containing the commonly used propellant HFA-134a has a global-warming potential equivalent to 0.13 kilograms of carbon dioxide³. The annual greenhouse-gas emissions from MDIs in the United Kingdom are equivalent to those from roughly 600,000 diesel cars.

Over the past two years, multiple agencies in the United Kingdom, where 70% of inhalers used are MDIs, have recommended schemes to reduce the carbon footprint of inhalers by switching to other types – particularly the propellant-free DPIs.

Wilkinson and his colleagues have calculated that switching half of all inhaler prescriptions in the United Kingdom to small-carbon-footprint devices, a target set by the UK Parliament’s Environmental Audit Committee, would save the equivalent of 288,000 tonnes of carbon dioxide every year⁴. That’s roughly equal to taking more than 61,000 cars off the road.

The target has led to a backlash from some patient advocate groups that say inhalers are necessary medication – not a lifestyle choice like eating meat or traveling by plane – and people might be unable to use alternatives

effectively, especially in an emergency. “We recognise the need to protect the environment, but it’s critically important that people with asthma receive the medicines they need to stay well and avoid a life-threatening asthma attack,” said Jessica Kirby, head of health advice at Asthma UK in London, in a statement responding to the study.

Some health-care professionals are cautious, too. “What concerns me is that uncritical implementation of this policy may lead to detriment to patient care,” Usmani says. “We may have struggled for many months or many years to stabilize a patient with asthma or COPD,” and changing their inhaler could put that at risk.

Meanwhile, companies are working to develop MDIs containing propellants with a smaller carbon footprint. Azapagic and her colleagues calculated that the global-warming impact of an inhaler containing one prominent alternative, HFA-152a, would be an order of magnitude smaller than those containing the common propellant HFC-134a.

In December, pharmaceutical company Chiesi in Parma, Italy, announced a €350-million (US\$385-million) effort to bring an MDI with a climate-friendly propellant to market by 2025. Pharmaceutical giant AstraZeneca announced in January that it plans to have a similar device ready by 2025.

Wilkinson argues that in some cases, moving away from MDIs could improve care for those with COPD. For example, DPIs might be more appropriate for people who have a tendency to inhale fast and hard (the correct technique for a DPI) or who have trouble coordinating their breath with an MDI. This could also be an opportunity to reduce over-prescription of inhaled corticosteroids, about 80% of which in the United Kingdom are provided in MDIs, but which are helpful for only a subset of people with COPD (see page S12). In any case, he says, the goal is not to change treatment plans that are working for individuals, but rather to encourage physicians to rethink their default prescribing practices.

“I don’t want patients to feel guilty about using MDIs,” Wilkinson says. “Ultimately, we need pharma companies to step up to the plate and sort out propellants that don’t have big carbon footprints. And it really looks like that’s happening now.”

Sarah DeWeerd is a science writer in Seattle, Washington.

1. Sanchis, J. et al. *Chest* **150**, 394–406 (2016).
2. Hanania, N. A. et al. *Chronic Obstr. Pulm. Dis.* **5**, 111–123 (2018).
3. Jeswani, H. K. & Azapagic, A. *J. Cleaner Prod.* **237**, 117733 (2019).
4. Wilkinson, A. J. K. et al. *BMJ Open* **9**, e028763 (2019).



DAVID GRAY/GETTY

Firefighters battle the bush fires that devastated Australia in 2019 and 2020.

Fireproofing the lungs

People with conditions such as COPD are vulnerable to wildfire pollution, but there is little advice on how to keep safe. **By Anna Nowogrodzki**

A few days into the new year, an older person came into John Hunter Hospital in Newcastle, Australia, wheezing and short of breath. Respiratory physician Peter Wark was on call at the time. He wasn't surprised to see someone with respiratory problems – Australia was enduring an unprecedented and devastating bush-fire season. Smoke from fires that had been raging kilometres away for the past four weeks had caused the air quality in the city to plummet.

Wark's patient already had chronic obstructive pulmonary disease (COPD), and her medical team had tried to prepare her for this kind of event. She had done her best to keep her windows and doors closed, despite a lack of air conditioning and some brutally hot days. And she had the anti-inflammatory drug

prednisone on hand to ease her symptoms. But still, she found breathing more and more difficult.

COPD is a common condition – it is the third leading global cause of death. And people with respiratory conditions such as COPD are some of the most vulnerable to particulate matter from air pollution and wildfires. Data from previous Australian bush fires, as well as wildfires in California, Colorado and North Carolina, show that people who have COPD visit the emergency department more frequently than usual during these events.

Yet physicians don't have the evidence they need to tell these vulnerable people what to do to protect their health. And researchers don't know what the effect of this exposure will be for everybody in the long term. Data suggest that long-term exposure to air pollution

leads to faster lung-function decline even in people with otherwise healthy lungs. "Other parts of the world, I think, should be watching very closely," says Wark, particularly the wildfire-prone US west coast.

"I find it rather unsettling that there are all these unknown things," says Guy Marks, a respiratory and environmental epidemiologist at the University of New South Wales, Sydney. "The scale of the fire that we've just had is unprecedented. It represents to me a clear turning point in our experience of the consequences of climate change."

Vulnerable lungs

Wark's patient improved just by being in the air-conditioned hospital. "We really didn't do anything else," he says. She was one of three or four older people with lung disease whom Wark remembers arriving at the hospital over the course of a few days. But he strongly suspects that many more people with respiratory diseases were suffering. "The ones who make it to the hospital are the tip of the iceberg," he says.

Wildfires are not good news for people with COPD. A 2019 review found evidence across multiple studies that visits to emergency departments increase for people with COPD¹. However, the data on hospitalization were mixed. Some studies found an association

between wildfire smoke and hospitalizations overall, and others did not, says Colleen Reid, a health geographer at the University of Colorado Boulder and an author of the review.

The most concerning pollutant for those who find themselves downwind of a wildfire is fine particulate matter less than 2.5 microns in diameter, says Reid. These PM_{2.5} particles are about four times smaller than a grain of pollen. In a 2019 paper, Reid found that PM_{2.5} levels increased sixfold downwind of a wildfire, whereas levels of ozone – another pollutant that can harm the lungs – increased less than twofold².

PM_{2.5} travels farther into the lung tissue than larger particles, almost reaching the tiny grape-like sacs called alveoli where gas exchange happens, says Nicholas Kenyon, a pulmonologist at University of California, Davis. He says that *in vitro* experiments suggest that, once in the lung tissue, the particles exacerbate chronic bronchitis (inflammation of the airways) and disrupt the layer of epithelial cells that line the airways.

It isn't clear exactly which chemicals in PM_{2.5} affect lung tissue, says Reid. "There could be a different chemical composition of the smoke depending on what's being burned," she says. Various studies have implicated different sets of chemicals in lung problems.

Scientists also don't know enough about the health impacts of ozone produced during wildfires, Reid says. Ozone causes airway inflammation and the formation of very unstable and highly reactive molecules. These free radicals can kill the lungs' epithelial cells, stripping the airways and leaving the lung tissue more vulnerable to viruses or allergens, says Kenyon. Research has found that higher ozone levels are correlated with increased hospital admissions and emergency-department visits for people with COPD³.

Uncertain response

There isn't enough research into how to protect the health of people with lung conditions from the spiralling effects of climate change, says Rupa Basu, an epidemiologist in the Office of Environmental Health Hazard Assessment at the California Environmental Protection Agency in Oakland. "Sometimes people look at all respiratory disease, which may not be the best way," she says. Lumping conditions together misses any differences in how people with, for example, COPD, asthma or cystic fibrosis are affected.

During wildfires, public-health officials often tell people to shelter in place, but there is limited research on how this affects people's health, Reid says. The benefit "really depends on where that place is", she says.

Poorly maintained properties and older homes tend to be leaky and let in smoke even with windows and doors closed, she explains. And people without air conditioning often leave their windows open. A study of indoor air quality during the 2016 and 2017 wildfire seasons in Denver, Colorado, found that most of the 28 low-income homes studied kept a window open for more than 12 hours a day, which more than doubled the levels of some pollutants in their homes⁴.

"There could be a different chemical composition of the smoke depending on what's being burned."

If people do shelter in place, evidence suggests that air purifiers such as high-efficiency particulate air (HEPA) filters in the home decrease particulate-pollution levels, says Reid. It is less clear whether purifiers improve the respiratory health of people with COPD. One study of 35 people with COPD found that HEPA filters had no effect on respiratory symptoms when used for 6 weeks⁵. Still, the reduction in particulates in the home is reason enough for many clinicians to recommend purifiers – especially to people who live very close to busy roads or notice soot on their windows. "I encourage them to get air purifiers," says Mary Rice, a pulmonary and critical care physician at Beth Israel Deaconess Medical Center in Boston, Massachusetts.

The costs can add up quickly, however, putting HEPA filters out of reach for people with low incomes. Air purifiers cost US\$100–600 in the United States. When used continuously, HEPA filters (costing \$90–175 each) need to be replaced every three months and use about \$30–90 of electricity per year – although running a purifier only during wildfires would cost less. And each room requires its own purifier.

When the smoke pollution is particularly bad, many people use N95 particulate respirator masks. These fit tighter than surgical masks and are designed to keep out particles as small as 0.3 microns – more than eight times smaller than PM_{2.5}. But many health professionals are concerned that people don't wear the masks properly or they don't fit well, and can therefore give people a false sense of security. "They can be helpful if they are put on properly on an individual that they fit correctly," says Reid. But, she explains, the masks don't fit well on people with facial hair, children or adults with smaller-than-average heads. Many people don't get a professional to test the fit of the

mask to ensure that it filters out particles as it is supposed to. Marks says there's not enough evidence to say whether N95 masks are beneficial for people with COPD. One study found that, for 14 people with mild COPD, wearing either an N95 mask or a mask that covered half the face affected breathing – in particular by limiting how quickly the person could exhale⁶.

Although it's not clear how people with COPD can protect themselves from the effects of smoke particles, medical interventions can help them if their symptoms worsen and breathing becomes more difficult. Ipratropium bromide and β -adrenergic agonists such as salbutamol can be taken to widen the bronchi. And prednisone – taken orally to reduce inflammation – helps some people. A 2003 study of people who had recently been discharged from hospital following an exacerbation of COPD showed that those who took prednisone for 5 days were considerably less likely to visit an emergency department within 30 days than were those who did not take it⁷. However, these strategies have not been tested specifically in people with acute severe smoke exposure in controlled trials, says Wark.

If needed, oxygen therapy and antibiotics can be provided in hospital, so it is also advised that people with COPD who live in at-risk locations have an action plan for getting to a medical centre during a wildfire.

Wildfires are short-term events, but climate change is already increasing their frequency, meaning that people in fire-prone areas will probably be exposed to wildfire smoke more often. There are few studies of the long-term health impacts of repeated wildfire-smoke exposure on either healthy people or those with COPD. But more generally, Rice says that long-term exposure to air pollution allowable within the current US Environmental Protection Agency standards "is associated with more rapid decline in lung function".

"I find myself rather frustrated at not having the answers," says Marks. As a COPD researcher, he says, "I get frequently asked, 'What are the risks and what should we do to protect ourselves?' And I give more or less the same answer: that we don't really know."

Anna Nowogrodzki is a journalist based in Boston, Massachusetts.

1. Reid, C. E. & Maestas, M. M. *Curr. Opin. Pulm. Med.* **25**, 179–187 (2019).
2. Reid, C. E. et al. *Env. Inter.* **129**, 291–298 (2019).
3. Malig, B. J. et al. *Environ. Health Perspect.* **124**, 745–753 (2016).
4. Shrestha, P. M. et al. *Int. J. Environ. Res. Public Health* **16**, 3535 (2019).
5. Blagev, D., Bride, D., Mendoza, D. & Horne, B. *Eur. Resp. J.* **54**, PA4454 (2019).
6. Harber, P. et al. *J. Occup. Environ. Med.* **52**, 155–162 (2010).
7. Aaron, S. D. et al. *N. Engl. J. Med.* **348**, 2618–2625 (2003).

María Victorina López Varela: Turning the tide

In 2002, pneumologist María Victorina López Varela at the University of the Republic and Hospital Maciel in Montevideo, Uruguay, joined forces with researchers in five other Latin American countries to launch the first large-scale assessment of the prevalence of chronic obstructive pulmonary disease (COPD) in the region — the PLATINO study.

What was the picture of COPD in Latin America in 2002?

There was no data on COPD prevalence in the region. Scientists, including myself, had worked on the disease for many years in universities, yet we didn't know the reality of COPD on our continent. That's why we launched the PLATINO study.

What did PLATINO involve?

It looked at the prevalence of COPD in around 5,000 people aged 40 or older, and the risk factors associated with the condition. It ran from 2002 to 2004 in five cities: Mexico City, Santiago, Montevideo, Caracas and São Paulo in Brazil. These are each the largest metropolitan areas in their respective countries, with a combined population of around 50 million. Their residents represent a variety of ethnic groups, and the cities are at different altitudes and in different climates. For the first time, we could find out how many people had the disease.

What did the study reveal?

It showed that COPD was a bigger problem in Latin America than previously thought. The overall crude prevalence of COPD was 14.3%, more than double the rate in the United States. But there were notable differences between the cities — the prevalence in Montevideo was 19.7%, whereas in Mexico City it was only 7.8% (M. Montes de Oca *et al.* *BRN Rev.* **3**, 3–17; 2017). When adjusted for key risk factors such as age, the prevalence ranking was largely maintained. We also learnt that 89% of people with COPD had previously gone undiagnosed — a higher rate than in other parts of the world.

How has the situation changed since then?

Not much. We performed a follow-up to PLATINO between 2008 and 2012 in the three cities with the highest prevalence:



María Victorina López Varela is trying to work out why the rate of COPD is so high in Uruguay.

Montevideo, Santiago (16.9%) and São Paulo (15.8%). We found the prevalence of stage 2–4 COPD to be quite stable, which suggests there is still a need for governments to take action.

Why was the prevalence in Montevideo in 2005 almost double the global rate?

We don't really know. A high smoking prevalence might have played a part, but smoking has declined since 2006, when anti-tobacco laws were introduced, and COPD prevalence has not. We also thought altitude might be a factor — we saw a correlation between the altitude of the five cities and their COPD rates (Mexico City sits at 2,240 metres above sea level, and Montevideo at only 35 metres). But a survey in five cities in Colombia did not find a correlation (A. Caballero *et al.* *Chest* **133**, 343–349; 2008).

There might also be a genetic explanation; we took blood samples to analyse, but need more funding to do so. And it could be that the way spirometry is commonly performed in Uruguay increases the likelihood of seeing signs of airway obstruction — people are often asked to blow for longer than the standard six seconds, according to our research.

What actions should be taken to improve the management of COPD in Latin America?

Definitely smoking cessation. However, 20% of those diagnosed with COPD are non-smokers.

Reducing exposure to indoor air pollution from biomass cooking stoves is also important. And, in the past few years, several studies have revealed other risk factors that should be addressed, such as experiencing respiratory disease as a child and being born to a severely malnourished mother. And of course, we need to continue to collect more data, especially on mortality.

Are you confident that steps to tackle COPD in Latin America will be taken?

We still need to raise awareness. Even now, much of the general population and many policymakers are unaware of COPD. Underdiagnosis is still above 80% in the region. To help solve this problem we developed the 'PUMA' questionnaire, which physicians can use to quickly identify people who need to be tested with spirometry. It won't solve everything by itself, however — we also need spirometry to be more widely available in primary care in Latin America.

It is also important to make sure that treatment is available to those who are diagnosed. The key to this is for the World Health Organization to maintain an up-to-date list of essential medicines for COPD.

Interview by Laura Vargas-Parada

This interview has been edited for length and clarity.

Boehringer Ingelheim: Reshaping the course of chronic obstructive pulmonary disease



AUTHORS

Abhya Gupta and Alberto de la Hoz

ADDRESS

Boehringer Ingelheim International GmbH,
Binger Strasse 173, 55216 Ingelheim am Rhein, Germany

The heritage of Boehringer Ingelheim in developing medicines for respiratory diseases spans almost 100 years. Within the respiratory disease area, innovative treatments for chronic obstructive pulmonary disease (COPD), asthma, idiopathic pulmonary fibrosis, lung cancer, allergic rhinitis and infantile respiratory distress syndrome have originated from our research facilities in Germany and around the world. Specifically for COPD, our products have been on the market for more than 40 years. Looking ahead, our aim is to significantly reshape the course of the disease.

WHY FOCUS ON COPD? BURDEN AND CHALLENGES

Affecting around 300 million people globally and representing a leading cause of death worldwide^{1,2}, COPD remains an area of significant clinical interest. The burden of the disease is expected to increase, fuelled by an ageing population² and increases in other risk factors for COPD^{3,4}. Although the main risk factor is cigarette smoking, air pollution and other environmental factors also play a part^{3,4}. The use of e-cigarettes (vaping) is also rising and, although its impact has not yet been elucidated, vaping may cause lung damage

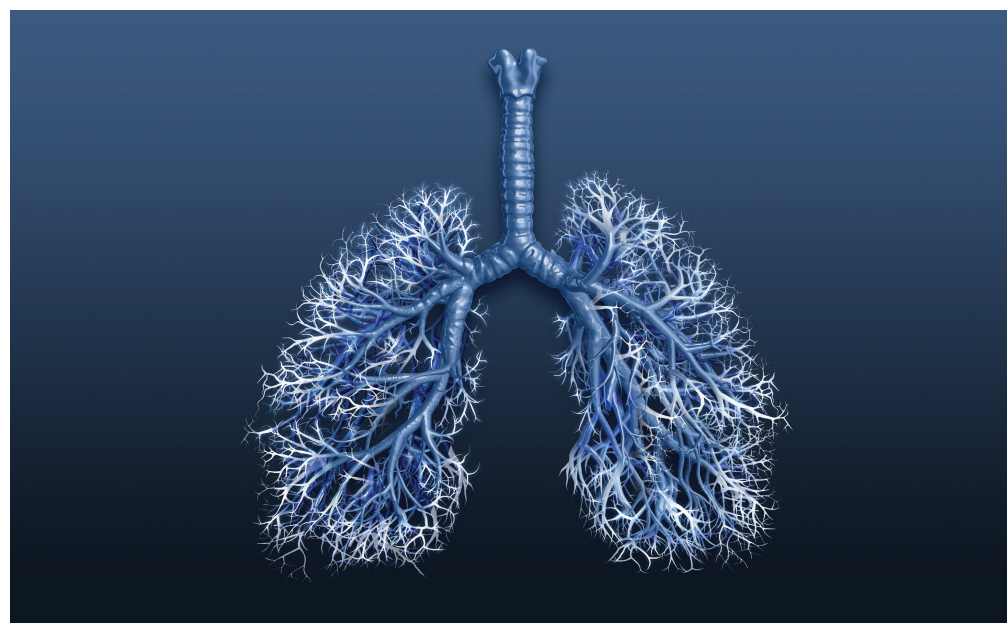


Figure 1. The small airways: The silent zone of the lungs.

and has been associated with reports of lung disease³.

COPD is a complex and heterogeneous incurable disease that can vary greatly in how it presents from patient to patient, even in those with a similar degree of airflow limitation⁵. It comes in several forms, one of which is emphysema, which causes progressive damage to the air sacs or alveoli of the lungs. Typical symptoms associated with underlying lung damage include breathlessness

(particularly during activity), a persistent cough and/or wheezing, and frequent bronchitis. Despite treatment, the breathing problems tend to get worse over time, and patients may become increasingly less active.

The early stages of the disease are largely asymptomatic; as a result, COPD is often diagnosed late, with more than three-quarters of individuals already having moderate or severe/very severe airflow obstruction

at the time of diagnosis⁶. The small airways, often referred to as the silent zone of the lungs, are regarded as one of the first sites where COPD develops. The surface area of the small airways is vast (**Fig.1**), which means it is possible for a large proportion of functional lung tissue to be destroyed before the disease becomes symptomatic.

Under-reporting of symptoms is common in COPD³. Many people without significant spirometric evidence

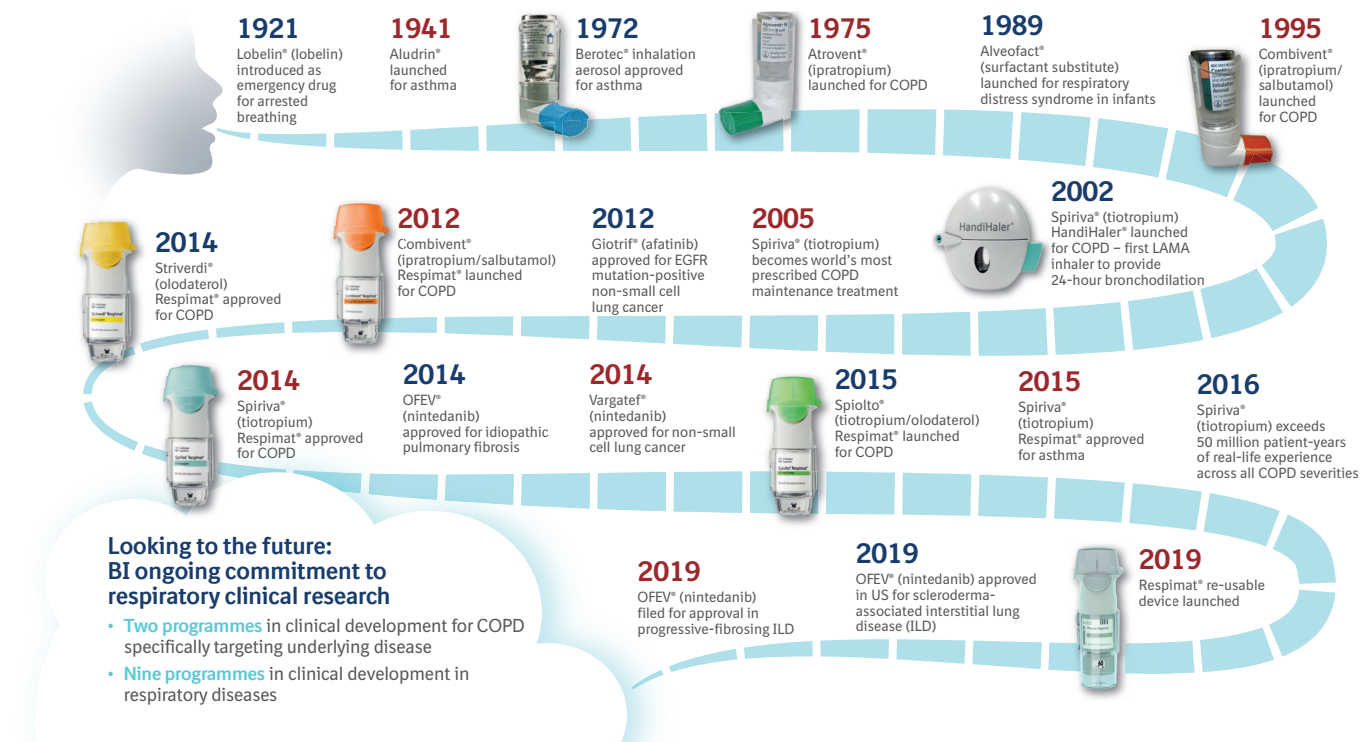


Figure 2. Boehringer Ingelheim: A heritage in respiratory diseases.

of COPD still suffer from respiratory symptoms such as breathlessness, cough and sputum production, and are at increased risk of death and disease progression. Nevertheless, these people are not on the physicians' radars according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) diagnostic criteria^{3,7}. This has prompted the proposal of new, broader criteria for diagnosing COPD, encompassing environmental exposure, clinical symptoms and computed tomography (CT) imaging, as well as spirometry findings⁷. Such an approach could potentially allow us to target the disease ahead of irreparable damage.

The goals of COPD treatment include reducing symptoms, preventing and treating exacerbations, reducing deaths related to the disease³, and changing the underlying nature of the disease (otherwise known as 'disease modification'). Current therapies provide symptomatic

relief and can reduce the risk or severity of exacerbations, but no pharmacological treatments are available yet to prevent the progression of lung destruction.

THE ROLE OF BRONCHODILATION

Bronchodilator therapy forms the mainstay of treatment for people with COPD. Bronchodilators work by altering the airways' smooth muscle tone, leading to widening of the airways, which, in turn, improves expiratory flow³. Since obstruction of the airways and associated symptoms (in particular, breathlessness) are primary concerns for most patients with COPD, the development of effective and well-tolerated bronchodilators has long been an important goal in COPD therapy.

As an alternative to increasing the dose of one bronchodilator, global guidance from GOLD recommends combining two bronchodilators with complementary mechanisms of action to increase the

impact of bronchodilation³. Indeed, combining a long-acting muscarinic antagonist (LAMA) with a long-acting β_2 -agonist (LABA) does improve bronchodilation and patient outcomes compared with a LAMA or LABA alone³.

We have an established history in the field of bronchodilators for use as maintenance therapy in people with COPD (Fig. 2). This includes the LAMA Spiriva® (tiotropium; first available in 2002), delivered using the HandiHaler® and then through the RespiMat® Soft Mist™ Inhaler, and the LABA Striverdi® (olodaterol; first available in 2014), delivered using the RespiMat®. More recently, in 2015, the fixed-dose dual LAMA/LABA combination therapy Spiolto®/Stiolto® (tiotropium/olodaterol), delivered using the RespiMat®, was approved for use in COPD and is currently available in 87 countries.

The combination of tiotropium/olodaterol increases lung function, improves health-

related quality of life and reduces breathlessness compared with tiotropium or olodaterol alone, with no observed difference in tolerability^{8,9}. Tiotropium/olodaterol also increases inspiratory capacity and exercise endurance¹⁰ versus placebo and single bronchodilator therapies. This is of particular importance because many patients can fall into a vicious cycle of activity avoidance, muscle deconditioning and increased breathlessness, leading to further physical inactivity to the detriment of their health-related quality of life. The use of bronchodilator therapies may allow individuals with COPD to maintain or increase their activity levels and improve their long-term prognosis.

THE CASE FOR EARLY ACTION

Even in people with mild COPD, symptoms such as breathlessness are often apparent and are generally associated with exercise intolerance and restriction of

physical activity. However, those who are affected may not seek medical help until symptoms are persistent and significant respiratory impairment is present.

For most patients, GOLD recommends initial pharmacological treatment with a single bronchodilator therapy, whereas dual LAMA/LABA therapy is recommended for more symptomatic patients³. However, there is a growing case for maximising bronchodilation as early as possible in the disease course. We have analysed data from our large clinical trial programme to evaluate the merits of dual bronchodilation. These results show the benefits of dual therapy with tiotropium/olodaterol in a wide variety of patient types, including treatment-naïve patients and patients with different degrees of symptoms and COPD severity⁹. Triple therapy with LAMA/LABA/inhaled corticosteroids also has an important place – escalation to triple therapy may be necessary for patients who develop further exacerbations on dual bronchodilation, especially among patients with higher eosinophil counts³.

INNOVATION IN INHALER DEVELOPMENT

How the drug treatments are delivered to their site of action in the lungs is hugely important. Historically, we were responsible for launching the first commercially available metered-dose inhaler (MDI) in Europe and the United States. Devices we have developed for use in people with COPD include the Atrovent® and Combivent® unit dose vials (for use in nebulisers), the HandiHaler® dry powder inhaler (DPI), the Atrovent® and Berodual® pressurised MDIs (pMDIs) and, most recently, the Respimat® Soft Mist™ Inhaler, which is available in

both disposable and re-usable formats (Fig. 2).

RESPIMAT® SOFT MIST™ INHALER: A FEAT OF ENGINEERING

Prior to the development of Respimat®, we were faced with a paradox. Although a large number of patients are prescribed inhalers such as DPIs, which require active inhalation for successful drug delivery, many are unable to inhale strongly enough or do not consistently perform a forceful inhalation manoeuvre, meaning that they do not use their inhalers effectively^{11,12}. pMDIs, on the other hand, have different challenges, only needing a relatively slow and deep inhalation, but requiring coordination between activating the device and breathing in to ensure that the drug particles reach the lung efficiently^{11,12}. Ineffective use of both DPIs and pMDIs can result in deposition of drug particles in the throat and mouth^{12,13}. It is essential, therefore, that an appropriate inhaler is matched to each patient, with device choice being important, as is the drug selected for treatment. In fact, GOLD recommends that patients who are unable to master their inhaler may need to consider a change in delivery device³.

In response to this conundrum, our in-house engineers developed the Respimat® Soft Mist™ inhaler (Fig. 3). This innovative new-generation propellant-free inhaler is driven by the desire to solve patients' problems with existing inhalers while maintaining a low carbon footprint^{12,14}. A key technical breakthrough in the development of the Respimat® was the Uniblock – the nozzle system of the inhaler – that combines filters and nozzles made of silicone and glass, inclined at a precise angle, so that two fine jets of liquid converge at a carefully controlled angle to create a slow-moving aerosol, from which the term 'soft mist' is derived, for optimised drug delivery¹². Drug solution is forced through this system using mechanical energy from a unique spring-loading mechanism to generate a fine aerosol of inhalable fine droplets¹². Respimat® generates a slow-moving, long-lasting 'soft mist' of drug, which can make it easier to inhale, and help provide a higher deposition of drug to the lungs compared with DPIs or pMDIs^{13,15}.

The use of the Respimat® inhaler also has good implications for treating small airways disease (SAD), an early site of lung deterioration in COPD^{13,16}. Believed to be present in around three-quarters of

patients with COPD, SAD is difficult to diagnose and assess due to the inaccessibility of the small airways^{16,17}. Given the overwhelming evidence for the importance of SAD in the development of COPD, the use of inhaled treatments that optimise delivery to the small airways is essential.

Importantly, Respimat® produces aerosol droplets of an appropriate size to ensure drug delivery throughout the lungs, including the small airways, without loss of small droplets during exhalation¹³. *In vivo* scintigraphy and *in vitro* models, based on CT imaging and computational fluid dynamics, have been used to demonstrate the effective deep lung deposition with the Respimat® inhaler; when compared with a range of different DPIs, Respimat® was shown to cause the lowest deposition of aerosol particles in the throat and the highest deposition in all regions of the lungs¹³.

CARBON FOOTPRINT AND ENVIRONMENTAL BENEFITS OF THE RESPIMAT® INHALER

Protecting the environment, conserving natural resources and promoting environmental awareness are principles that are highly valued at Boehringer Ingelheim.

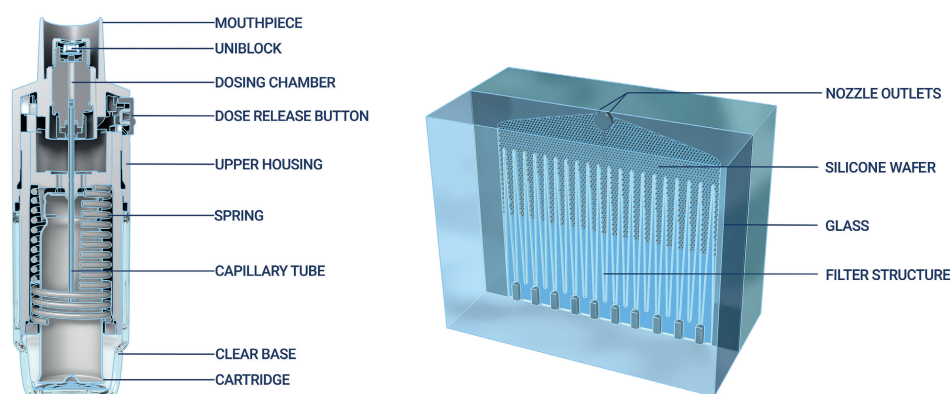


Figure 3. A cross-sectional image of the Respimat device¹². This work is licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>).

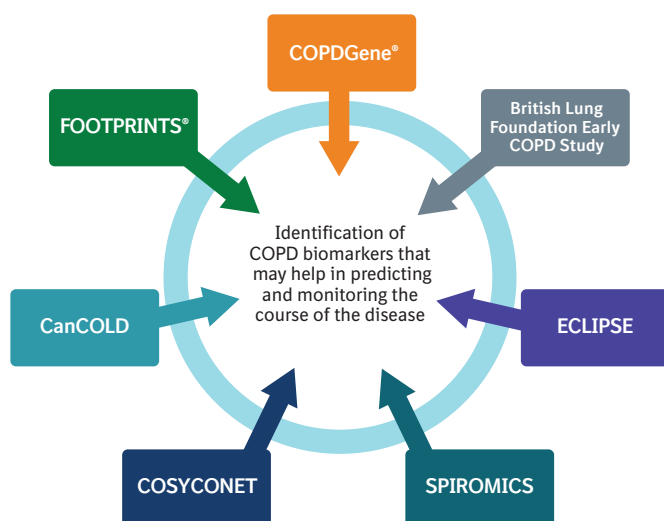


Figure 4. Key studies investigating biomarkers in chronic obstructive pulmonary disease (COPD).

A vast number of people use inhalers; in fact, in the United Kingdom (UK) an estimated 1.2 million people are living with diagnosed COPD (most of whom will be prescribed inhaled therapy), while 5.4 million people are currently receiving treatment, again mostly inhaled therapy, for asthma. Since MDIs account for nearly 4% of UK National Health Service (NHS) greenhouse gas emissions, use of hydrofluoroalkane propellants in inhalers is coming under increasing scrutiny. The NHS has pledged to set a target for at least 50% of prescribed inhalers to have low global warming potential values by 2022¹⁴. Respimat® uses mechanical energy and is propellant-free, making it an environmentally friendly option. In fact, the product carbon footprint of the original disposable Respimat® inhaler is approximately 20 times smaller than hydrofluoroalkane pMDI products and similar to that of DPIs¹⁴.

Further reduction in environmental impact has been provided by the novel second-generation Respimat® re-usable inhaler, which was first launched in 2019. Prompted by feedback from patients

and physicians, this was an evolution of the disposable Respimat®, and an example of our commitment to patient-centric and environmentally friendly inhaler design. The new re-usable inhaler can be used with up to six cartridges, providing increased convenience for patients while also improving ease of use¹⁵. Crucially, over the six-month lifespan of the re-usable Respimat®, there is a threefold reduction in its carbon footprint compared with the disposable Respimat® inhaler, representing a considerable contribution to sustainability¹⁴. In recognition of this contribution, the Respimat® re-usable was the winner of the 2020 Pharmapack Eco-Design Award. The Pharmapack awards celebrate the latest innovations from packaging companies within the drugs, medical devices, health products and veterinary drugs sectors.

FUTURE DIRECTIONS: THE ROLE OF PRECISION MEDICINE IN COPD

We are building upon our strong heritage in bronchodilator and inhaler development with

ongoing clinical development programmes aimed at improving the outlook for people with COPD. COPD is a heterogeneous and progressive disease. Both COPD itself and emphysema – a major cause of airflow limitation in COPD – are associated with reduced and declining lung function. Although current pharmacological treatment options can improve lung function, ultimately, they do little to prevent or reverse the decline in forced expiratory volume in 1 second – the most commonly used marker of disease severity and progression in COPD – over time. A more in-depth understanding of disease progression is needed to target future therapies.

Biomarkers, including those associated with emphysema and SAD, may assist in characterising patients, both in terms of response to therapy and also in predicting and monitoring the course of disease. They may also help to identify mechanisms responsible for lung destruction in COPD, which in turn could identify new treatment targets that bronchodilators – although a mainstay of therapy – do not directly impact.

To supplement findings from ongoing studies such as COPDGene® (copdgene.org), the British Lung Foundation Early COPD Study (imperial.ac.uk/blf-early-copd-partnership), ECLIPSE (eclipse-copd.com), SPIROMICS (spiromics.org), COSYCONET (www-mhh.asconet.net) and CanCOLD (cancold.ca), we are carrying out the FOOTPRINTS® study (NCT02719184), which hopes to identify biomarkers of COPD disease progression – particularly emphysema or lung function loss progression – over a three-year period (Fig. 4).

FOOTPRINTS® is assessing a range of biomarkers in blood and other biofluids in former smokers with COPD versus otherwise

healthy former smokers. These include biomarkers relating to protease activity, extracellular matrix biomarkers and inflammatory biomarkers. Airway wall thickness, emphysema and air trapping are being assessed using CT.

By correlating patient characteristics, biomarkers of inflammation and tissue integrity, imaging findings and lung function, the progression of lung destruction and emphysema progression may be better understood. This in turn may support the development of future treatments to slow or halt disease progression in COPD and identify those patients who are at greatest risk for disease progression. FOOTPRINTS® is also aiming to identify different COPD patient types and their relative risks of disease progression. This could include the identification of ‘rapid progressors’, who have greater unmet need.

FUTURE DIRECTIONS: MODIFYING THE COURSE OF THE DISEASE

In line with our clinical research in precision medicine, we are also developing innovative treatments that can further improve the lives of people with COPD. We are researching new molecules with the potential to slow lung destruction and therefore modify the course of the disease, with the ultimate goal of developing a maintenance treatment that slows emphysema progression. By targeting patients with preserved ratio-impaired spirometry (a high-risk group for developing COPD), it might be possible in the future to even prevent or reverse lung damage⁷.

Recently, attention has focused on the link between COPD-related inflammation and tissue damage, and the imbalance between a type of enzymes called serine

proteinases and their inhibitors (Fig. 5). Several serine proteases, including neutrophil elastase, cathepsin G and proteinase-3, are implicated in the destruction of alveolar tissue¹⁸, and are therefore potential targets for treatment.

Although these mechanisms have been investigated previously^{19,20}, the degree of inhibition on neutrophil elastase activity in patients was only moderate. We are currently evaluating neutrophil elastase and cathepsin C inhibition in preclinical and clinical studies to explore the exciting potential for these molecules in COPD and other respiratory diseases, with the hope of moving beyond improving lung function into modifying the course of the disease.

Although there is much to celebrate regarding our heritage in respiratory diseases – spanning almost 100 years of innovation in drug and device development – there is continuing medical need for therapies that can change the underlying nature of the disease. Given that currently available treatments focus on symptom control and risk reduction, we need to maximise the effectiveness of those therapies by focusing on early and accurate diagnosis, optimising bronchodilation and tailoring the treatment selection to the patient needs. This in turn could help to optimally manage symptoms in order to maintain and improve patient quality of life and reduce the risk of experiencing exacerbations. It is also important to match the inhaler selection to patients' ability and preference, including the use of innovative inhalers such as the Respimat® and Respimat® re-usable, to help optimise management of COPD. Meanwhile, our ongoing research and clinical programmes aim to target the power of precision medicine.

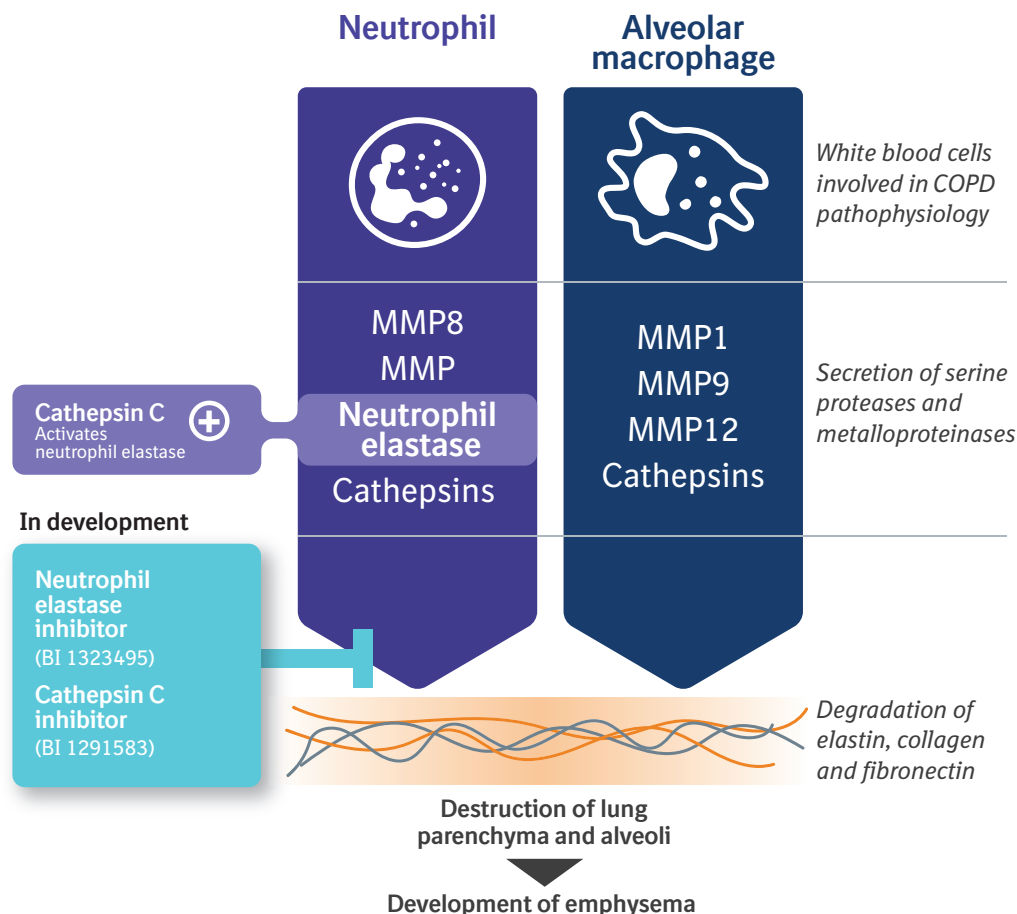


Figure 5. Potential targets to modify the course of chronic obstructive pulmonary disease (COPD).

By enhancing our knowledge of COPD-specific biomarkers, and continuing to assess innovative new treatment targets, we hope to meet our goal of modifying the course of the disease.

Only by doing this can we improve the outlook for people with COPD.

ACKNOWLEDGEMENTS

The authors would like to thank Cindy Macpherson from MediTech Media for her help in the development of this paper.

REFERENCES

- James, S. L. et al. *Lancet* **392**, 1789–1858 (2018).
- Mathers, C. D. & Loncar, D. *PLoS Medicine* **3**, e442 (2006).

- Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease (2020 report). <https://goldcopd.org/wp-content/uploads/2019/11/GOLD-2020-REPORT-ver1.0wms.pdf>. Published 2019. Accessed February 14, 2020.
- Jiang, X.-Q., Mei, X.-D. & Feng, D. *J. Thorac. Dis.* **8**, E31–E40 (2015).
- Agusti, A. et al. *Respir. Res.* **11**, 122 (2010).
- Mapel, D. W., Dalal, A. A., Blanchette, C. M., Petersen, H. & Ferguson, G. T. *Int. J. Chron. Obstruct. Pulm. Dis.* **6**, 573–581 (2011).
- Lowe, K. E. et al. *Chron. Obstr. Pulm. Dis.* **6**, 384–399 (2019).
- Buhl, R. et al. *Eur. Respir. J.* **45**, 969–979 (2015).
- Singh, D. et al. *Respir. Res.* **17**, 73 (2016).
- O'Donnell, D. E. et al. *Eur. Respir. J.* **49**, 1601348 (2017).

- Melani, A. S. et al. *Respir. Med.* **105**, 930–938 (2011).
- Wachtel, H., Kattenbeck, S., Dunne, S. & Disse, B. *Pulm. Ther.* **3**, 19–30 (2017).
- Ciciliani, A. M., Langguth, P. & Wachtel, H. *Int. J. Chron. Obstruct. Pulm. Dis.* **12**, 1565–1577 (2017).
- Hänsel, M., Bambach, T. & Wachtel, H. *Adv. Ther.* **36**, 2487–2492 (2019).
- Dhand, R., Eicher, J., Hansel, M., Jost, I., Meisenheimer, M. & Wachtel, H. *Int. J. Chron. Obstr. Pulm. Dis.* **14**, 509–523 (2019).
- McNulty, W. & Usmani, O. S. *Eur. Clin. Respir. J.* **1**, 25898 (2014).
- Crisafulli, E. et al. *Respiration* **93**, 32–41 (2017).
- Pandey, K. C., De, S. & Mishra, P. K. *Front. Pharmacol.* **8**, 512 (2017).
- Stockley, R. et al. *Respir. Med.* **107**, 524–533 (2013).
- Watz, H. et al. *Pulm. Pharmacol. Ther.* **56**, 86–93 (2019).